

Threshold estimation based on a P -value framework

Bodhisattva Sen, Moulinath Banerjee and George Michailidis

Columbia University, University of Michigan and University of Michigan

August 26, 2010

Abstract

We use p -values as a discrepancy criterion for identifying the threshold value at which a regression function takes off from its baseline value – a problem that is motivated by applications in omics experiments, systems engineering, pharmacological dose-response studies and astronomy. In this paper, we study the problem in a controlled sampling setting, where multiple responses, discrete or continuous, can be obtained at a number of different covariate-levels. Our procedure involves testing the hypothesis that the regression function is at its baseline at each covariate value using the sampled responses at that value and then computing the p -value of the test. An estimate of the threshold is provided by fitting a stump, i.e., a piecewise constant function with a single jump discontinuity, to the observed p -values, since the corresponding p -values behave in markedly different ways on different sides of the threshold. The estimate is shown to be consistent, as both the number of covariate values and the number of responses sampled at each value become large, and its finite sample properties are studied through an extensive simulation study. Our approach is computationally simple and can also be used to estimate the baseline value of the regression function. The procedure is illustrated on two motivating real data applications. Extensions to multiple thresholds are also briefly investigated.

Keywords: baseline value, change point, consistent estimate, controlled sampling, misspecified model, stump function.

1 Introduction and Problem Formulation

In a number of applications, the data follow a regression model where the regression function μ is constant at its baseline value τ_0 up to a certain covariate threshold d^0 and stays above τ_0 at higher covariate levels. For example, consider the data in the left panel of Figure 1 that depict the average delay of customers as a function of the loading of a complex queueing system (more details about the system are given in Section 5.3). It can be seen that for small loadings the delay is rather small, while a clear positive trend emerges for larger loadings. The system's operator is interested in identifying this threshold with high precision, as well as the level of the baseline τ_0 , since such knowledge specifies an operational regime of low average delay, whose value can be announced to potential customers. An example from a different scientific domain is shown in the right panel of Figure 1. It depicts the expression levels of a gene over time obtained from multiple cell-lines, where again the function stays at its baseline level up to some time, then rises sharply and subsequently flattens out (more details about the underlying experiment are given in Section 5.3). This problem requires procedures that can handle multiple change-points of the regression function, namely where it deviates from the baseline value and also where it starts flattening out towards the end of the range of the covariate. These thresholds are of interest as they represent important stages of progression of the cell from normalcy to malignancy.

Problems with identical structure also arise in pharmacological dose-response studies, where

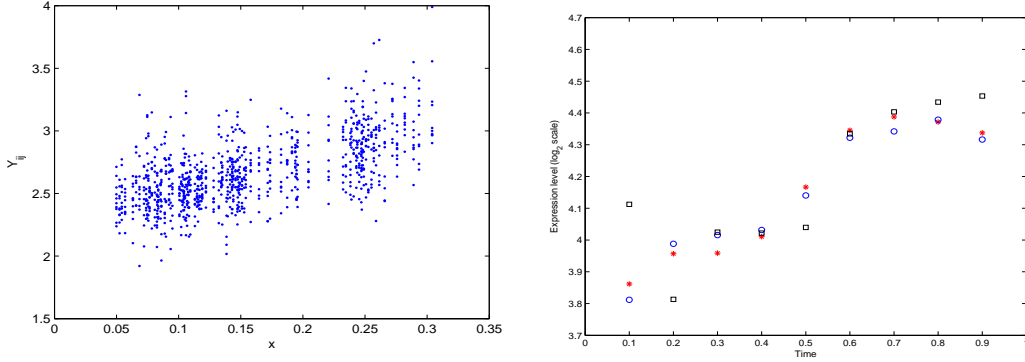


Figure 1: Left panel: Data of delay versus loading from a complex queueing system. Right panel: Gene expression levels over time.

$\mu(x)$ provides information about reaction to dose-level x and is typically at the baseline value up to a certain dose, referred to in the literature as the minimum effective dose (MED); see Chen and Chang (2007) and Tamhane and Logan (2002) and the references therein. Similar problems arise in toxicological applications; see, for example, Cox (1987), who uses parametrically specified threshold models. Yet another field of application is astronomy; one of particular interest to the authors deals with estimating the “tidal” radius of a dwarf spheroidal galaxy (see Sen et al. (2009)). The mean velocity of the stars along the major axis of the dwarf spheroidal galaxy, as a function of the distance from the center, is assumed to be constant to the left of the “tidal” radius (threshold parameter), and takes off from this baseline value, due to interactions with the gravitational field of the Milky Way, as we move to the right of the threshold. Further applications and extensions are discussed in Section 6. We also note that our current problem of interest is a special and important case of the more general question of identifying the region where a function (defined on a general Euclidean space) assumes an extremal (minimum/maximum) value.

Formally, we consider a non-negative regression function $\mu(x)$ on $[0, 1]$ with the property that $\mu(x) = \tau_0$ for $x \leq d^0$ and $\mu(x) > \tau_0$ for $x > d^0$ for some $d^0 \in (0, 1)$. As already mentioned, quantities of prime interest are d^0 and τ_0 that need to be estimated from noisy data $\{Y_i, X_i\}_{i=1}^n$, with X_i ’s assuming values in $[0, 1]$ and $Y_i = \mu(X_i) + \epsilon_i$, where ϵ_i is a mean 0 error. We call d^0 the “ τ_0 threshold” of the function μ . In this generality, i.e., without any assumptions on the behavior of the function in a neighborhood of d^0 , the estimation of the threshold d^0 is a hard problem and has not been extensively addressed in the literature. In the simplest possible setting of the problem posited, the regression function $\mu(x)$ has a jump discontinuity at d^0 . In this case, d^0 corresponds to a change-point for μ and the problem reduces to estimating this change-point in a regression model. Such change-point models are very well studied; see, for example, Mueller (1992), Loader (1996), Mueller and Song (1997), Koul and Qian (2002), Lan, Banerjee and Michailidis (2009) and references therein.

The problem becomes significantly harder when μ is continuous at d^0 ; in particular, the smoother the regression function in a neighborhood of d^0 , the more challenging the estimation. For example, if d^0 is a cusp of μ of some *known* order p (i.e., the first $p - 1$ right derivatives of μ at d^0 equal 0 but the p -th does not, so that d^0 is a change-point in the p -th derivative), one can obtain nonparametric estimates for d^0 using either kernel based (Mueller (1992)) or wavelet based (Raimondo (1998)) methods. The convergence rate of such estimates decreases dramatically with p (see Raimondo (1998)), even with parametric specifications of μ to the right of the unknown d^0 (Feder (1975)). In fact, if μ is infinitely differentiable at d^0 , no estimate possibly converges at a polynomial rate, although we are not aware of any such

concrete result in the literature. However, in most applications, the degree of differentiability of μ at d^0 *will not be known*, which makes necessary the development of adaptive estimation procedures that do not require prior information about the smoothness of the underlying regression function.

In many applications of interest, μ is known to be continuously increasing, and a natural candidate for an adaptive procedure is isotonic regression; further, it automatically avoids the specification of bandwidths and equivalent smoothing parameters (see Robertson et. al. (1988) and Silvapulle and Sen (2005)). Applications of isotonic regression in calibration type problems involving thresholds are discussed, for example, in Osborne (1991), Gruet (1996) and Tang et al. (2010). It is known that the isotonic regression estimate $\hat{\mu}$ is piecewise constant, usually possessing a flat stretch of 0 for small values of the covariate. Thus, one can prescribe $\hat{d} \equiv \inf\{x : \hat{\mu}(x) > 0\}$ as an estimate of d^0 , but this, in most cases, severely underestimates the true threshold. It is possible to employ penalized isotonic estimates or replace the 0 in the definition of \hat{d} by a positive sequence η_n converging to 0 (at an appropriate rate), but hardly anything is known in the literature about the theoretical properties of such procedures. Moreover, in many applications the assumption of global monotonicity of μ may not hold.

This paper develops a novel approach for the *consistent estimation* of d^0 for situations where *multiple observations can be sampled* at a given covariate value that (i) does not require knowledge of the smoothness of μ at d^0 , (ii) does not require computing an explicit estimate of μ and (iii) is computationally simpler than most nonparametric procedures. Note that this multiple observations per “dose” setting is the scenario for both our data applications and also for most of the dose–response studies in pharmacological experiments.

Note that, in both the motivating applications introduced in the first paragraph, the experimenter can specify the values of the covariate (either the load of the queueing system or the time point when the cell-line is harvested) and subsequently obtain the corresponding sample responses (delays or expression levels). In general, these sampled responses are expensive to obtain. In the first example, this is the case since generating the responses involves a discrete event simulation, with the response at each loading obtained by averaging over a number of events (e.g., customers who have received complete service from the system). For large scale systems, this may require tens of thousands of such events, which may exceed the allotted budget of resources. An alternative strategy is to rely on responses obtained from a fairly “small” set of events for each selected loading, which would lead to significantly more noisy observations, as seen in the left panel Figure 1. In the second example, carrying out the biological experiment is fairly costly primarily due to the labor involved (preparation of cell-lines, microarray hybridization and processing). The obtained data can be noisy due to the inherent biological variability of the cell-lines. Nevertheless, the goal is to identify the transition point(s) and the corresponding levels of the threshold from such noisy data. The developed nonparametric methodology allows us to resolve this issue in a satisfactory manner. Specifically, it relies on testing for the value of μ at design levels of the covariate. The obtained test statistics are then used to construct p -values which, under mild assumptions on μ , behave in markedly different fashions on either side of the threshold d^0 and it is this discrepancy that is used to construct an estimate of d^0 .

The remainder of the paper is organized as follows: in Section 2, the proposed procedure is introduced in the case of *known* τ_0 , its properties derived and some extensions discussed. The case of *unknown* τ_0 is examined in Section 3. Section 4 briefly investigates the generalization to multiple change points. The performance of the procedure based on simulated data is studied in Section 5 where comparisons with other competing methods are also presented. The procedure is also illustrated on real data from the two motivating applications. Some concluding remarks are drawn in Section 6, while proofs of most technical results are given

in the Appendix.

2 The p -value procedure: the known τ_0 case

To introduce and motivate the proposed p -value procedure, we first consider the case with τ_0 known. Specifically, let $Y = \mu(X) + \epsilon$, where μ is a function on $[0, 1]$ and

$$\mu(x) = \tau_0 \text{ for } x \leq d^0, \text{ and } \mu(x) > \tau_0 \text{ for } x > d^0, \quad (1)$$

for $d^0 \in (0, 1)$. *Note that no other assumptions are made on the behavior of μ around d^0 .* The covariate X is sampled from a Lebesgue density p_X on $[0, 1]$ and ϵ is independent of X and distributed as $N(0, \sigma^2)$, with σ known. Relaxations of some of these assumptions will be discussed later.

As argued above, estimation of d^0 via direct estimation of μ is a hard problem. The use of p -values allows a relatively easy solution when multiple responses can be sampled at each covariate value. More specifically, we have:

$$Y_{ij} = \mu(X_i) + \epsilon_{ij}, \quad i = 1, 2, \dots, n; \quad j = 1, 2, \dots, m, \quad (2)$$

with $N = m \times n$ being the total budget of samples. The ϵ_{ij} 's are i.i.d. and distributed like ϵ above and the X_i 's are i.i.d. from p_X . The assumption of equal number of replicates (m) at each X_i can actually be relaxed to a certain extent and is discussed briefly in the concluding discussion, but for the sake of ease of exposition of the key ideas we assume this throughout the paper.

At dose $X_i = x$, we test the null hypothesis $H_{0,x} : \mu(x) = \tau_0$ against the alternative $H_{1,x} : \mu(x) > \tau_0$ using the test statistic $T(x) = \frac{\sqrt{m}(\bar{Y}_{i,\cdot} - \tau_0)}{\sigma}$ where $\bar{Y}_{i,\cdot} = \frac{1}{m} \sum_{j=1}^m Y_{ij}$. The observed p -value for this test is $p^{(m)}(x) = 1 - \Phi(T(x))$, since $T(x)$ is distributed as $N(0, 1)$ under the null hypothesis. From the n different dose levels, we obtain n p -values $p^{(m)}(X_1), p^{(m)}(X_2), \dots, p^{(m)}(X_n)$.

Under the null hypothesis which holds to the left of d^0 , the p -values have a Uniform(0,1) distribution. To the right of d^0 , where the null hypothesis fails, the distributions of the p -values change and as m becomes large, the p -values converge to the degenerate value 0. This dichotomous behavior of the p -values on either side of d^0 can be used to prescribe *consistent* estimates of the latter. A natural way to capture this discrepancy, which we explore in this paper, is to consider the expected p -value curve at stage m , formally $\nu_m(x) \equiv E(p^{(m)}(x))$. Notice that this is identically 0.5 for all $x \leq d^0$, irrespective of m , while for $x > d^0$, it converges to 0 as m increases. We illustrate this in Figure 2 assuming $\sigma = 0.5$ for $m = 10, 20, 50$ and 100. This simple observation can be used to construct estimates of d^0 which do not involve estimating μ . We can fit a stump to the observed p -values, with levels 1/2 and 0 on either side of the break-point and prescribe the break-point of the best fitting stump (in the sense of least squares) as an estimate of d^0 . The virtue of this approach lies in the fact that we are able to estimate the threshold consistently, as established rigorously below, by merely fitting a simple mis-specified working model. Its success relies on the fact that the p -values eventually show stump like (dichotomous) behavior which no estimate of μ could have inherited, regardless of sample size. We describe our approach quantitatively below.

For convenience, denote $Z_{im} \equiv p^{(m)}(X_i)$. Letting $\xi_d(x) = \frac{1}{2}1(x \leq d)$, to find the stump with levels 1/2 and 0 that best approximates the observed p -values $Z_{im}, i = 1, 2, \dots, n$, we minimize

$$\mathbb{M}_{m,n}(d) = \sum_{i=1}^n \{Z_{im} - \xi_d(X_i)\}^2 = \sum_{i: X_i \leq d} \left(Z_{im} - \frac{1}{2}\right)^2 + \sum_{i: X_i > d} Z_{im}^2 \quad (3)$$

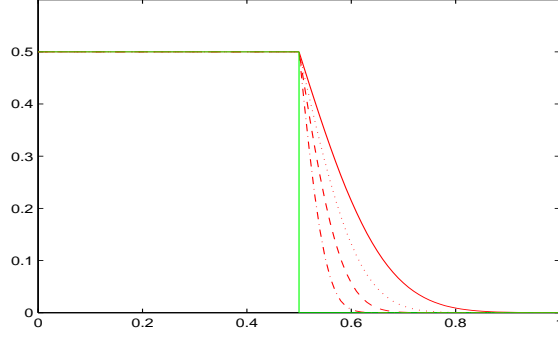


Figure 2: Expected p -value curves converging to a stump with increasing m .

over $d \in [0, 1]$. Let $\hat{d}_{m,n} = \arg \min_{d \in [0,1]} \mathbb{M}_{m,n}(d)$, which is a natural estimator of d^0 .

However, in practice, the use of $1/2$ and 0 as the stump levels may not always be the best strategy. Firstly, the p -values to the right of d^0 may not be small enough to be well approximated by 0 for a finite m ; secondly, we may often have situations where the Z_{im} 's are *not exact but approximate* p -values. For example, if σ was unknown in the above setting, $T(x)$ would take the form $\sqrt{m} \bar{Y}_i / \hat{\sigma}$, where $\hat{\sigma}$ is some estimate of σ , in which case the Z_{im} 's would not be uniformly distributed (or even close to a uniform for modest m). In such cases, one can adopt a more adaptive approach by allowing the stump levels to converge to $1/2$ and 0 with increasing m , or by keeping the stump-levels unspecified and estimating them from the data itself. Such adaptive procedures might provide a better fit to the observed p -values and improve the precision of the estimate of d^0 .

We summarize next the proposed estimation procedure and establish its consistency. Formally, the setup is as follows: Consider the (possibly heteroscedastic) regression model $Y = \mu(X) + \epsilon$ with μ as in (1), $E(\epsilon|X) = 0$, $\text{Var}(\epsilon|X) > 0$ and the covariate X following a Lebesgue density p_X on $[0, 1]$. The available data from this model $\{X_i, \{Y_{ij}\}_{j=1}^m\}_{i=1}^n$ are exactly as in (2). The steps of the procedure are the following:

1. For $i = 1, 2, \dots, n$, let $p^{(m)}(X_i) \equiv Z_{im}$ denote the observed (potentially approximate) p -value based on a test of the hypothesis $H_{0,i} : \mu(X_i) = 0$ against the alternative $H_{1,i} : \mu(X_i) > 0$, using data $\{Y_{ij} : j = 1, 2, \dots, m\}$, such that $Z_{1m}, Z_{2m}, \dots, Z_{nm}$ are i.i.d. as well.
2. Fit a stump $\alpha_m 1(x \leq d) + \beta_m 1(x > d)$ to $\{Z_{im}\}_{i=1}^n$, where α_m and β_m are known non-negative quantities that converge to $1/2$ and 0 respectively. For $d \in [0, 1]$ define:

$$\mathbb{M}_{m,n}(d) = \sum_{i: X_i \leq d} (Z_{im} - \alpha_m)^2 + \sum_{i: X_i > d} (Z_{im} - \beta_m)^2 \quad (4)$$

and let $\hat{d}_{m,n} = \arg \min_{d \in [0,1]} \mathbb{M}_{m,n}(d)$. We can choose $\alpha_m = 1/2$ and $\beta_m = 0$ for all m to get back to the setting of (3).

3. We can even let the data choose the optimal α_m and β_m by setting

$$\hat{\theta}_{m,n} \equiv (\hat{\alpha}_{m,n}, \hat{\beta}_{m,n}, \hat{d}_{m,n}) = \arg \min_{\theta \equiv (\alpha, \beta, d) \in [0,1]^3} \sum_{i=1}^n \{Z_{im} - \alpha 1(X_i \leq d) - \beta 1(X_i > d)\}^2.$$

Theorem 2.1 *Consider the above setup of the problem and let $\nu_m(x) = E(Z_{1m}|X_1 = x)$. Assume further that (a) $\nu_m(x) \rightarrow \nu(x) := (1/2) 1(x \leq d^0)$ for each x , as $m \rightarrow \infty$, and (b) $p_X(x) > \kappa > 0$ for $x \in [d^0 - l, d^0 + l]$ for some (small) $l > 0$. We then have:*

- (i) $\hat{d}_{m,n} \xrightarrow{P} d^0$ as $m, n \rightarrow \infty$, i.e., given $\epsilon, \eta > 0$, there exists a positive integer K , such that for $m, n \geq K$, $P\{|\hat{d}_{m,n} - d^0| > \epsilon\} < \eta$.
- (ii) $\hat{\theta}_{m,n} \xrightarrow{P} \theta_0 \equiv (1/2, 0, d^0)$ as $m, n \rightarrow \infty$.

The theorem is proved in the Appendix.

Flexible modeling of the p -value curve: In many applications the curve μ is an increasing continuous function, whence the expected p -value curves are continuous decreasing functions converging to a stump with increasing m as in Figure (2). One can then also use continuous parametric working models to take into account the shape of the p -value curve for finite m . Figure (2) suggests looking at sigmoidal curves. We propose and explore one such family, a 2-parameter sigmoidal family \mathcal{G} next:

$$\mathcal{G} = \left\{ \psi_{d,\alpha}(x) \equiv \frac{1}{2} \mathbf{1}\{x \leq d\} + \frac{e^{-\alpha(x-d)}}{1 + e^{-\alpha(x-d)}} \mathbf{1}\{x > d\} : d \in [0, 1], \alpha \geq 0 \right\}. \quad (5)$$

For $x \leq d$, $\psi_{d,\alpha}(x) = 1/2$ and this models the region where $\mu = 0$ (or a constant). For $x > d$, $\psi_{d,\alpha}$ is decreasing, thus mimicking the finite sample behavior of the expected p -value curve ν_m . It can be shown that consistent estimates of d^0 may be obtained with this misspecified class of models, as well. Consistent estimation of d^0 is possible because in the limit (as $\alpha \rightarrow \infty$ and $d \rightarrow d^0$) the sigmoid converges to the stump (indicator) function ν .

We estimate the parameters d and α of the working model by solving a least squares problem, i.e.,

$$(\tilde{d}_{m,n}, \tilde{\alpha}_{m,n}) = \arg \min_{d \in [0,1], \alpha \geq 0} \mathbb{G}_{m,n}(d, \alpha) \equiv \frac{1}{n} \sum_{i=1}^n \{Z_{im} - \psi_{d,\alpha}(X_i)\}^2. \quad (6)$$

It can be shown under assumptions (a) and (b) of Theorem 2.1 that $\tilde{d}_{m,n} \xrightarrow{P} d^0$, as $n, m \rightarrow \infty$, using arguments similar to those in the proof of that theorem.

It should be noted that there is nothing special about using the parametric model (5) to estimate d^0 . Any reasonable class of models that includes (or can converge to) the stump function would, in principle, yield consistent estimators of d^0 . However, we focus on the stump ξ_d , mostly because of the conceptual/computational simplicity and accurate finite sample performance (as illustrated in the simulation study in Section 5, in which results for the 2-parameter sigmoidal family are also included). Figure 3 shows the true expected p -value curve (shown by the solid curve) and illustrates the three methods of finding d^0 in a single simulation run with $\sigma = 0.5$, $d^0 = 0.5$, $n = 20$, $m = 10$: (1) fitting a stump with levels $1/2$ and 0 (shown by the solid vertical line denoting the estimated value of d^0); (2) fitting a stump with adaptive levels as in Theorem 2.1 (ii) (shown by the dashed horizontal lines) and the estimated d^0 (shown by the dotted vertical line); (3) the fitted sigmoid model defined in (6) (shown by the dashed-dotted curve).

2.1 The procedure under relaxed assumptions

While the assumptions of normality of errors and known variance were used to motivate the procedure of stump based approximations to the p -values in the previous section, Theorem 2.1 shows that these assumptions can be relaxed considerably. Further, as the theorem does not require the responses to be continuous, the procedure is also valid in discrete response settings, one of which we illustrate in the subsequent discussion. We discuss below a number of different scenarios that can arise in practice.

(i) *Homoscedastic errors with unknown variance:* Suppose that the ϵ_{ij} 's are continuous i.i.d.

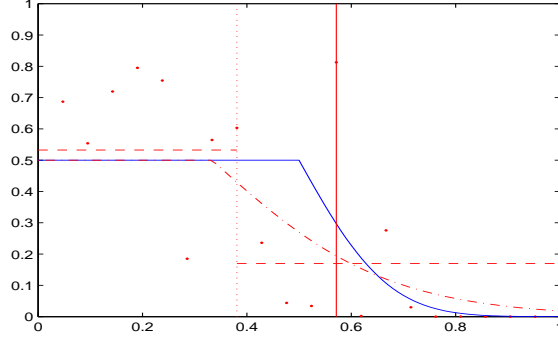


Figure 3: Fitting different models to the p -value curve.

random variables with mean 0 and unknown variance σ^2 and independent of $\{X_i\}_{i=1}^n$. First, define “working” p -values $p^{(m)}(X_1), p^{(m)}(X_2), \dots, p^{(m)}(X_n)$ as $p^{(m)}(X_i) = 1 - \Phi\left(\frac{\sqrt{m}(\bar{Y}_{i\cdot} - \tau_0)}{\hat{\sigma}(X_i)}\right)$ where $\hat{\sigma}^2(X_i) = \sum_{j=1}^m (Y_{ij} - \bar{Y}_{i\cdot})^2 / (m - 1)$. The p -values are then independent, $\nu_m(x) = E(p^{(m)}(X_1) | X_1 = x)$ converges to $\nu(x)$ and Theorem 2.1 applies, yielding the consistency of the stump-based least squares estimates of d^0 . However, owing to the homoscedasticity of the errors, the constructed p -values are clearly sub-optimal since we could have replaced $\hat{\sigma}(X_i)$ in the definition of the i ’th p -value by $\hat{\sigma}_{m,n}$, where $\hat{\sigma}_{m,n}^2 = \sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - \bar{Y}_{i\cdot})^2 / (mn - n)$ is the standard pooled estimate of σ^2 (and is significantly superior to each $\hat{\sigma}(X_i)$). Thus, least squares estimates of d^0 based on (b) and (c) in the build-up to Theorem 2.1, using such *improved* p -values would certainly yield consistent estimates of d^0 . Unfortunately, we are no longer quite in the setting of Theorem 2.1, as the presence of $\hat{\sigma}$ in each p -value makes them dependent, and therefore cannot invoke its conclusions to deduce consistency. We tackle the consistency issue for this problem below. Roughly speaking, as m, n grow, $\hat{\sigma}$ stabilizes quickly, the p -values become approximately independent and the scenario approaches that of Theorem 2.1.

Theorem 2.2 Assume that (a) $\inf_{d^0 + \beta \leq d \leq 1} \mu(d) > \tau_0$ for any $\beta > 0$, and that (b) $p_X(x) < K$ for $x \in [d^0, d^0 + \eta]$ for some $\eta > 0$. Let $\hat{d}_{m,n}$ be the least squares estimate of d^0 obtained as: $\hat{d}_{m,n} = \arg \min_d \left[\sum_{i: X_i \leq d} (Z_{im} - \frac{1}{2})^2 + \sum_{i: X_i > d} (Z_{im} - 0)^2 \right]$, where $Z_{im} = 1 - \Phi\left(\frac{\sqrt{m}(\bar{Y}_{i\cdot} - \tau_0)}{\hat{\sigma}_{m,n}}\right)$; these being the ‘improved’ p -values alluded to in (i) above. Then $\hat{d}_{m,n} \xrightarrow{p} d^0$ as $m, n \rightarrow \infty$.

A condensed version of the proof is available in the Appendix.

Remark: If we *know* that the errors are normal, the statistics $\sqrt{m}\bar{Y}_{i\cdot}/\hat{\sigma}$ follow a t_{mn-n} distribution. Letting $F_{t_{mn-n}}$ denote the corresponding distribution function, we can take $Z_{im} = 1 - F_{t_{mn-n}}\left(\frac{\sqrt{m}\bar{Y}_{i\cdot}}{\hat{\sigma}}\right)$ and the corresponding estimates of d^0 continue to be consistent.

(ii) *Heteroscedastic errors:* Consider a regression set-up with continuous responses and heteroscedastic errors: i.e. $\sigma^2(x) \equiv \text{Var}(\epsilon | X = x)$ varies with x . Thus, $\sigma^2(X_i)$ is estimated by $\hat{\sigma}^2(X_i) = \sum_{j=1}^m (Y_{ij} - \bar{Y}_{i\cdot})^2 / (m - 1)$ and $Z_{im} = 1 - \Phi\left(\frac{\sqrt{m}(\bar{Y}_{i\cdot} - \tau_0)}{\hat{\sigma}(X_i)}\right)$; in this case, pooling is no longer possible unlike case (i). It is not difficult to see that $\nu_m(x) = E(Z_{im} | X_i = x)$ converges to $\nu(x)$ as above and the conclusions of Theorem 2.1 hold. In case the errors are conditionally normally distributed, Φ in the definition of the p -values can be replaced by $F_{t_{m-1}}$, the distribution function of the t_{m-1} distribution.

(iii) *Discrete responses:* Settings where the response is discrete can also be considered. One can think of the covariate values as dose-levels and suppose, at each dose level, X_i , that a binomial experiment with m independent subjects is performed, and the response Y_i is the number that show a reaction to the dose. The function $\mu(x)$ is the probability that a subject yields a reaction at dose x and is assumed to be at a baseline value $p_0 > 0$ for $x \leq d^0$ and greater than p_0 otherwise (note that $p_0 = 0$ gives a pathological situation).

We base our p -value at dose level X_i on the normalized statistic $(Y_i - m p_0) / \sqrt{m p_0 (1 - p_0)}$ with (working p -value) $Z_{im} = 1 - \Phi((Y_i - m p_0) / \sqrt{m p_0 (1 - p_0)})$. Alternatively, the Z_{im} 's could also be defined to be p -values based on the exact binomial distribution: i.e., $Z_{im} = 1 - F_{m,p_0}(Y_i)$ where F_{m,p_0} is the distribution function of $\text{Binomial}(m, p_0)$. The conditions of Theorem 2.1 are easy to verify in either case and the conclusions of the theorem continue to hold.

2.2 A Digression: Composite Hypotheses

Consider now a situation where $\mu(x)$ is known to be strictly less than (a known) ζ_0 if $x < d^0$ and strictly greater than ζ_0 for $x > d^0$. We keep the behavior at d^0 unspecified. For a monotone function μ this reduces to estimating its inverse at ζ_0 . The latter problem has been well-studied in the literature (Banerjee and Wellner (2005), Tang et. al. (2010) and references therein) but only under explicit shape and/or smoothness constraints on μ . Our formulation, on the other hand, is much broader in scope as it requires neither monotonicity, nor smoothness assumptions.

To formulate the problem, consider the model posited in (2) with homoscedastic normal errors and μ as in the beginning of the previous paragraph. At dose $X_i = x$ the statistic $T(x) = \sqrt{m}(\bar{Y}_i - \zeta_0) / \sigma$ is used to test the composite hypothesis, $H_{0,x} : \mu(x) < \zeta_0$ versus $H_{1,x} : \mu(x) > \zeta_0$, with rejection for large values of $T(x)$. Construct an approximate p -value as $p^{(m)}(x) = 1 - \Phi(T(x))$. Define $\nu_m(x) = E[p^{(m)}(x)]$. It is easy to check that $\nu_m(x)$ converges to 1 as $m \rightarrow \infty$ for $x < d^0$ and to 0 for $x > d^0$. This suggests

$$\hat{d}_{m,n} = \arg \min_d \mathbb{M}_{m,n}(d) \equiv \sum_{i: X_i \leq d} (Z_{im} - 1)^2 + \sum_{i: X_i > d} Z_{im}^2$$

as a natural estimate of d^0 , with $Z_{im} \equiv p^{(m)}(X_i)$, $i = 1, 2, \dots, n$.

We present an analogue of Theorem 2.1 (in the current setting) that generalizes the above strategy. We start by summarizing the proposed estimation procedure.

1. Identical to Step (a) in the build-up to Theorem 2.1, except for $p^{(m)}(X_i) \equiv Z_{im}$ denoting the observed (potentially approximate) p -value based on a test of the hypothesis $H_{0,i} : \mu(X_i) < \zeta_0$ against the alternative $H_{1,i} : \mu(X_i) > \zeta_0$.
2. Step (b) remains the same as before but with α_m converging to 1.
3. Step (c) remains unaltered.

Theorem 2.3 *Let $\nu_m(x) = E(Z_{1m} | X_1 = x)$ and further assume that (a) $\nu_m(x) \rightarrow 1$ for each $x < d^0$ and $\nu_m(x) \rightarrow 0$ for each $x > d^0$, as $m \rightarrow \infty$, and, (b) $p_X(x) > \kappa > 0$ for $x \in [d^0 - l, d^0 + l]$ for some (small) $l > 0$. Then, we have*

1. $\hat{d}_{m,n} \xrightarrow{P} d^0$ as $m, n \rightarrow \infty$.
2. $\hat{\theta}_{m,n} \xrightarrow{P} \theta_0 \equiv (1, 0, d^0)$ as $m, n \rightarrow \infty$.

The proof of the theorem is similar to that of Theorem 2.1 and is therefore skipped.

3 The Case of an Unknown τ_0

In this section, we address the more realistic situation where τ_0 is unknown. As will become apparent in subsequent developments, a number of complications arise in this setting. In order to focus on the main ideas, we confine ourselves to the setting of continuous responses. Throughout this section, we consider the problem of fitting a one-parameter stump with levels $1/2$ and 0 on either side of the jump-location, the main parameter of interest.

Homoscedastic errors: Consider first, the homoscedastic error setting as in (i) of Section 2.1. The consistency of the least squares estimate of d^0 in this model by fitting a stump (with levels $1/2$ and 0 on either side of the jump) to the observed p -values was established in Theorem 2.2. However, with an unknown τ_0 , these p -values can not be constructed. A natural alternative is to replace τ_0 with some appropriately consistent estimate. The next theorem shows the consistency of the estimate of d^0 obtained by fitting a stump-model to the observed p -values when an appropriate estimate of τ is used. For $\sigma, \tau > 0$, define:

$$Z_{im}^\sigma(\tau) = 1 - \Phi(\sqrt{m}(\bar{Y}_i - \tau)/\sigma) \quad (7)$$

for $i = 1, 2, \dots, n$.

Theorem 3.1 *Consider model (2) where the errors $\{\epsilon_{ij}\}$ are i.i.d with variance σ_0^2 and independent of $\{X_i\}$. Suppose that $\tilde{\tau} \equiv \tilde{\tau}_{m,n}$ is a consistent estimator of τ_0 such that $\sqrt{m}(\tilde{\tau} - \tau_0) = o_p(1)$. Let $\tilde{\sigma} \equiv \tilde{\sigma}_{m,n}$ be a consistent estimator of σ_0 and let $\tilde{d} \equiv \tilde{d}_{m,n}$ be estimated as:*

$$\tilde{d}_{m,n} = \arg \min_{d \in [0,1]} \left[\sum_{i: X_i \leq d} \left\{ Z_{im}^{\tilde{\sigma}_{m,n}}(\tilde{\tau}_{m,n}) - \frac{1}{2} \right\}^2 + \sum_{i: X_i > d} \left\{ Z_{im}^{\tilde{\sigma}_{m,n}}(\tilde{\tau}_{m,n}) - 0 \right\}^2 \right]. \quad (8)$$

Then $\tilde{d}_{m,n} \xrightarrow{P} d^0$ as $m, n \rightarrow \infty$.

Remark: In a situation where d^0 may be safely assumed to be greater than some known positive η , an estimate of τ_0 satisfying the condition of the above theorem can be obtained by taking the average of the response values on the interval $[0, \eta]$. However, this does not offer a satisfactory solution, since such an η may not be known in various applications. Also, even if η is known but small, the estimate thus obtained will be unsatisfactory unless n is really large. Below, we adopt a more principled approach to the estimation of τ_0 that does not require such background knowledge, once again using p -values.

We now focus on constructing an explicit estimator $\tilde{\tau}$ of τ_0 as required in Theorem 3.1, using p -values. Recall the definition of $Z_{im}^\sigma(\tau)$ in (7). Let $\tau > \tau_0$ and note that as m increases, for $\mu(X_i) < \tau$, $Z_{im}^{\hat{\sigma}_{m,n}}(\tau)$ goes to 1 in probability, while for $\mu(X_i) > \tau$, $Z_{im}^{\hat{\sigma}_{m,n}}(\tau)$ goes to 0 in probability. For any $\tau < \tau_0$, it is easy to see that $Z_{im}^{\hat{\sigma}_{m,n}}(\tau)$ always goes to 0 in probability, whereas when $\tau = \tau_0$, $Z_{im}^{\hat{\sigma}_{m,n}}(\tau)$ goes to 0 for $X_i > d^0$, but is uniformly distributed on $(0, 1)$ for $X_i < d^0$ for every m . Thus, it is only when $\tau = \tau_0$ that $Z_{im}^{\hat{\sigma}_{m,n}}(\tau)$'s are the closest to $1/2$ for a substantial number of i 's. This suggests a natural estimate for τ_0 : namely,

$$\hat{\tau}_{m,n} = \arg \min_{\tau} \sum_{i=1}^n \{Z_{im}^{\hat{\sigma}_{m,n}}(\tau) - 1/2\}^2. \quad (9)$$

Once $\hat{\tau}_{m,n}$ is obtained, an estimate of d^0 , say $\hat{d}_{m,n}$ can be obtained by taking $\tilde{\tau}_{m,n}$ to be $\hat{\tau}_{m,n}$ in (8) and $\tilde{\sigma}_{m,n}$ to be $\hat{\sigma}_{m,n}$. This method of estimating d^0 and τ_0 is referred to, subsequently, as *Method 1*. Theorem 3.2 shows that under some mild conditions on the function μ , $\sqrt{m}(\hat{\tau}_{m,n} - \tau_0)$ is $o_p(1)$. This along with the fact that $\hat{\sigma}_{m,n}$ is consistent for σ_0 , implies that $\hat{d}_{m,n}$ is consistent for d^0 by Theorem 3.1.

Theorem 3.2 Consider the same setup as in Theorem 3.1. Further suppose that the regression function μ satisfies:

(A) Given $\eta > 0$, there exists $\epsilon > 0$ such that, for every $\tau > \tau_0$,

$$\int_{\{x > d^0 : |\mu(x) - \tau| \leq \epsilon\}} p_X(x) dx < \eta.$$

Also assume that ϕ_m , the density function of $\sqrt{m}\bar{\epsilon}_1/\sigma_0$, converges pointwise to ϕ , the standard normal density. Then $\sqrt{m}(\hat{\tau}_{m,n} - \tau_0) \xrightarrow{P} 0$ as $m, n \rightarrow \infty$.

Remark: Condition (A) is guaranteed if, for example, μ is strictly increasing to the right of d^0 although they hold under weaker assumptions on μ . In particular it rules out flat stretches to the right of d^0 . Note that the assumption that ϕ_m converges to ϕ is not artificial, since convergence of the corresponding distribution functions to the cdf of the standard normal is guaranteed by the central limit theorem.

An alternative method (*Method 2*): Notice that the previous method involves the estimation of d^0 in two steps: first by estimating τ_0 and subsequently using this estimate to approximate the p -values to which a stump is fitted, as described in (8). An alternative *one-step* method for estimating d^0 (that avoids estimating τ_0) is presented, when μ is increasing. Define $\xi_n(x) = E(Y|X \leq x)$. A natural estimate of $\xi_n(x)$ is given by $\hat{\xi}_n(x) = \sum_{i: X_i \leq x} \sum_{j=1}^m Y_{ij} / \{m \sum 1(X_i \leq x)\}$. It can be easily checked that for $x \in (0, 1)$, $\hat{\xi}_n(x) - \xi(x)$ is $O_p((mn)^{-1/2})$, a fact that will be used later. As our estimate of d^0 we propose:

$$\tilde{d}_{m,n} = \arg \min_{d \in [0,1]} \left[\sum_{i: X_i \leq d} \{Z_{im}^{\hat{\sigma}_{m,n}}(\hat{\xi}_n(X_i)) - 1/2\}^2 + \sum_{i: X_i > d} \{Z_{im}^{\hat{\sigma}_{m,n}}(\hat{\xi}_n(X_i)) - 0\}^2 \right]. \quad (10)$$

We do not formally establish the consistency of this procedure in the Appendix, but provide a heuristic discussion below. The performance of this method is also assessed via simulation studies. Once $\tilde{d}_{m,n}$ has been obtained, an estimate of τ_0 is given by $\hat{\xi}_n(\tilde{d}_{m,n})$.

Discussion: Denote the quantity within the big square brackets on the right side of the above display by $\Psi(d)$. Let $\eta > 0$ be such that $0 < d^0 - \eta < d^0 + \eta < 1$. Consider the difference $\Psi(d^0 - \eta) - \Psi(d^0)$, which can be written as:

$$\sum_{d^0 - \eta < X_i \leq d^0} \left[\{Z_{im}^{\hat{\sigma}_{m,n}}(\hat{\xi}_n(X_i))\}^2 - \{Z_{im}^{\hat{\sigma}_{m,n}}(\hat{\xi}_n(X_i)) - 0.5\}^2 \right].$$

Now, for any $X_i \in (d^0 - \eta, d^0]$,

$$Z_{im}^{\hat{\sigma}_{m,n}}(\hat{\xi}_n(X_i)) = 1 - \Phi \left(\frac{\sqrt{m}(\bar{Y}_{im} - \tau_0)}{\hat{\sigma}_{m,n}} + \frac{\sqrt{m}(\tau_0 - \hat{\xi}_n(X_i))}{\hat{\sigma}_{m,n}} \right).$$

For sufficiently large m, n , the first term within the brackets on the right side of the above display is approximately distributed like a standard normal, while the second term is small by virtue of the fact that $\sqrt{m}(\hat{\xi}_n(X_i) - \tau_0)$ is $O_p(n^{-1/2})$ (where we tacitly make use of the fact that these X_i 's are all bounded away from 0). It follows that the right side is approximately distributed like a Uniform(0,1) denoted by U_i . Thus, $\Psi(d^0 - \eta) - \Psi(d^0)$ behaves approximately like $\sum_{i: d^0 - \eta < X_i \leq d^0} U_i^2 - \sum_{i: d^0 - \eta < X_i \leq d^0} (U_i - 0.5)^2$ for U_i 's that are approximately uniform and weakly correlated for sufficiently large m, n . But this quantity will tend to be non-negative with high probability. A similar argument can be used to show that $\Psi(d^0 + \eta) - \Psi(d^0)$ will tend to be non-negative with high probability, when μ is increasing, which we leave to the reader. This illustrates why the minimizer of Ψ is close to d^0 , with high probability, in the

long run.

Heteroscedastic Errors: We briefly discuss the case of heteroscedastic errors as in (ii) of Subsection 2.1. Let $\hat{\sigma}_m^2(X_i) \equiv \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2 / (m-1)$ and $Z_{im}^{\hat{\sigma}_m(X_i)}(\tau) = 1 - \Phi(\sqrt{m}(\bar{Y}_i - \tau) / \hat{\sigma}_m(X_i))$. To construct a consistent estimator $\hat{\tau}$ of τ_0 we can use (9) with $Z_{im}^{\hat{\sigma}_m(X_i)}(\tau)$ changed to $Z_{im}^{\hat{\sigma}_m(X_i)}(\tau)$. Now, Method 1 can be implemented to obtain an estimator for d^0 using (8) with $Z_{im}^{\hat{\sigma}_m(X_i)}(\hat{\tau}_{m,n})$ replaced by $Z_{im}^{\hat{\sigma}_m(X_i)}(\hat{\tau})$. Method 2 can also be implemented by replacing the superscript $\hat{\sigma}_{m,n}$ by $\hat{\sigma}_m(X_i)$ in (10).

4 Multiple Change Points

While our procedure was primarily motivated by applications with single baseline thresholds, it can be extended without much difficulty to the case of multiple thresholds. To illustrate the idea, consider a regression model where the function $\mu(x)$, with x varying in $(0, 1)$ is at its baseline value, say τ_0 , on an interval of the form $[a, b]$ with $0 < a < b < 1$ and stays above the baseline elsewhere. For ease of illustration, we restrict ourselves to the situation with a continuous response and homoscedastic errors with unknown variance, as in Section 2.1 (i). As in that problem, we would construct p -values at each point, $\{Z_{im}\}_{i=1}^n$. Our estimates of a and b would be obtained as:

$$(\hat{a}_{m,n}, \hat{b}_{m,n}) = \arg \min_{a < b} \left[\sum_{X_i \in [a,b]} (Z_{im} - 1/2)^2 + \sum_{X_i \notin [a,b]} Z_{im}^2 \right].$$

The computation of the minimizer would proceed by searching over all pairs $(X_{(i)}, X_{(j)})$ with $i < j$. By extending the techniques used in our proofs one can establish consistency of the above least squares estimates without much additional difficulty. In case τ_0 is unknown, one can use exactly the same estimate of τ_0 as advocated in (9). Similar reasoning as in the discussion preceding (9) shows that $\hat{\tau}_{m,n}$ would be consistent for τ_0 in this setting. It is not difficult to see how the procedure would extend to other kinds of regions, for example, a baseline zone which consists of a disjoint union of finitely many intervals, though the computational complexity increases rapidly with the number of disjoint intervals concerned. The structure of the baseline zone will, of course, be determined by the application under consideration.

The procedure can also be extended to allow for situations where μ attains its minimum and maximum values, say τ_{min} and τ_{max} , on disjoint intervals $[a, b]$ and $[c, d]$, as is the case for the gene expression data. Assume, for simplicity, that the minimum and maximum values are known (or that very reliable estimates are available). One then applies the procedure in the previous paragraph to determine $[a, b]$. To determine $[c, d]$, note that in the model $-Y_i = -\mu(x) - \epsilon_i$, $[c, d]$ is the region on which $-\mu$ hits its minimum and therefore the above procedure can once again be applied with the signs of the responses flipped.

When the minimum and the maximum are unknown, a natural temptation would be to estimate the minimum via (9) in the original problem, and the maximum via (9) again in the sign-flipped problem, separately; however, that becomes suspect in this situation since μ has two flat stretches while the method advocated in (9) is theoretically justified only when flat stretches at levels larger than the minimum are ruled out; see Assumption (A) of Theorem 3.2 in this context. However, in many situations, there will be a reasonable degree of separation between the minimum and the maximum and it will be possible to identify these values up to mutually disjoint intervals (for more on this issue see the analysis of the gene data example). In such cases, the procedure in (9) can again be brought into play by doing *restricted searches* over the τ domain: to identify the minimum, minimize the criterion function in (9) *only over the interval in which the minimum is expected to lie* (as opposed to searching over the entire putative range of μ) and do a similar analysis for the maximum (by switching to

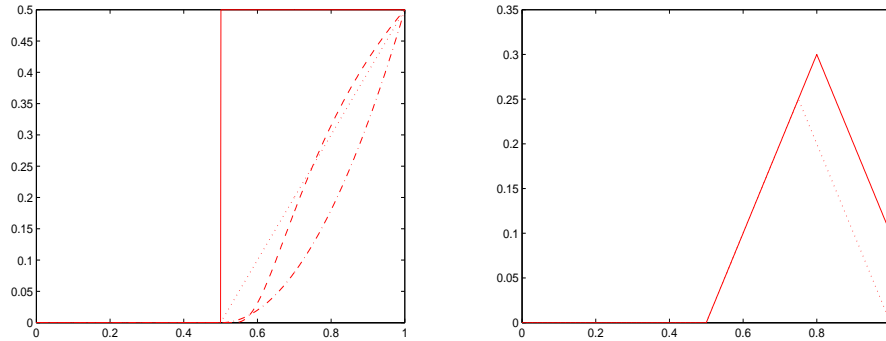


Figure 4: Plots of the four increasing regression functions μ 's (left panel) and the two tent shaped μ 's (right panel).

the sign-flipped problem). In the absence of other flat stretches (besides the minimum and the maximum) this procedure will estimate the extremal values accurately.

5 Simulation Results and Data Analysis

We first study the performance of our proposed methods through an extensive simulation study. We then compare our approach with other competing procedures mostly developed in the dose-response setting. The methodology is subsequently illustrated on the motivating examples of the complex queueing system and the gene expression data. In the simulation study, we compare in a number of settings, the simplest model of the one-parameter stump to both the more complex, sigmoid and the three-parameter adaptive stump models. We undertake a comprehensive evaluation of the two methods proposed for estimating the threshold d^0 in the presence of unknown τ_0 , we investigate the performance of the proposed methodology when the threshold is located close to the boundary of the design space and finally we discuss how one should allocate a fixed budget of samples between number of doses n and replicates m .

5.1 Simulation Studies

In our numerical studies, six choices of the regression function μ are considered. Four of these are non-decreasing and the remaining two are “tent”-shaped. The four monotone ones are depicted in the left-panel of Figure 4, all of which are 0 to the left of $d^0 = 0.5$. Specifically, M_0 , shown by the solid line, is a stump and is identically equal to 0.5 to the right of d^0 ; M_1 described by the dotted line is a piece-wise linear function (a kink-model) rising from 0 to 0.5 between d^0 and 1; M_2 , the convex curve, grows like a quadratic beyond d^0 , whilst $M_3(x) = \exp(-\lambda/(x - 0.5))1(x > 0.5)$ (for an appropriate λ so that $M_3(1) = 0.5$) is infinitely differentiable at d^0 . Thus, from M_0 to M_3 we have four functions exhibiting increasing smoothness at d^0 .

Our two “tent”-shaped regression functions are depicted in the right panel of Figure 4; they are identically 0 till $d^0 = 0.5$ and tent-shaped beyond: M_4 rises linearly from 0.5 to 0.75 with unit slope and then declines symmetrically from 0.75 to 1, reaching 0 at the point 1, while M_5 is a slight variant of M_4 rising linearly with unit slope from 0.5 till 0.8 and then decreasing with unit slope from 0.8 till 1.0. Estimation of $d^0 \equiv 0.5$ for the tent-shaped functions is expected to be more challenging than for the monotone ones. Note that for the monotone curves, the “signal” keeps on increasing as we move to the right of d^0 , whereas with the tent-shaped curves, the signal starts to weaken beyond a point, and with M_4 especially,

Table 1: RMSEs for the one-parameter stump (first entry of each column) and the two-parameter sigmoid (second entry of each column) with $\sigma = 0.1$ (top) and 0.3 (bottom panel) for five models and different choices of m and n .

$\sigma = 0.1$					
(m, n)	M_0	M_1	M_2	M_3	M_4
(5, 5)	0.153, 0.193	0.154, 0.195	0.163, 0.217	0.154, 0.199	0.154, 0.195
(5, 10)	0.130, 0.170	0.130, 0.169	0.119, 0.164	0.099, 0.150	0.116, 0.174
(10, 10)	0.127, 0.169	0.131, 0.171	0.110, 0.159	0.093, 0.142	0.124, 0.170
(10, 20)	0.079, 0.126	0.068, 0.120	0.097, 0.126	0.075, 0.110	0.060, 0.120
(10, 50)	0.030, 0.066	0.033, 0.069	0.106, 0.100	0.082, 0.082	0.033, 0.073
(20, 50)	0.028, 0.062	0.030, 0.066	0.088, 0.088	0.073, 0.077	0.030, 0.071
(50, 100)	0.014, 0.033	0.014, 0.032	0.073, 0.058	0.071, 0.061	0.014, 0.033
$\sigma = 0.3$					
(m, n)	M_0	M_1	M_2	M_3	M_4
(5, 5)	0.154, 0.195	0.155, 0.202	0.201, 0.235	0.171, 0.219	0.208, 0.255
(5, 10)	0.130, 0.169	0.132, 0.175	0.200, 0.222	0.140, 0.178	0.214, 0.258
(10, 10)	0.130, 0.164	0.137, 0.175	0.166, 0.198	0.119, 0.163	0.148, 0.226
(10, 20)	0.069, 0.115	0.089, 0.134	0.177, 0.181	0.116, 0.135	0.120, 0.220
(10, 50)	0.032, 0.062	0.086, 0.093	0.193, 0.153	0.128, 0.114	0.095, 0.212
(20, 50)	0.032, 0.062	0.061, 0.079	0.162, 0.134	0.117, 0.092	0.061, 0.135
(50, 100)	0.014, 0.033	0.039, 0.041	0.131, 0.093	0.098, 0.081	0.039, 0.050

comes back to the baseline value at the right boundary.

For each allocation pair (m, n) and regression function $\mu(x)$, we generate responses $\{Y_{i1}, Y_{i2}, \dots, Y_{im}\}$ at covariate value $x_i = i/(n+1)$, for $i = 1, 2, \dots, n$, with $Y_{ij} = \mu(x_i) + \epsilon_{ij}$, the $\{\epsilon_{ij}\}$'s being i.i.d. $N(0, \sigma^2)$. Two different choices of σ (0.1 and 0.3) are considered. For each combination of model, allocation and noise-level (σ), the performance, in terms of Root Mean Square Error (henceforth RMSE), of the estimator under consideration is evaluated based on 2000 replicates. Note that a non-random uniform design is used to generate the covariates, which is a slightly different procedure from generating the n covariates uniformly from $(0, 1)$. Of course, asymptotically, it does not make a difference, but for small n , the regularity of the uniform grid has a salutary effect on the performance of the least squares estimates.

One-parameter stump vs sigmoid model: Table 1 gives the RMSEs for estimating d^0 , for a number of different (m, n) combinations, models M_0 through M_4 and two values for σ (0.1 and 0.3), using a one-parameter stump $(1/2)1(x \leq d)$ and a two-parameter sigmoid, as described in (5). The first key observation is that in the relatively high “signal-to-noise” regime ($\sigma = 0.1$), the inference problem is relatively easier and the one-parameter stump outperforms the two-parameter sigmoid almost uniformly (with an occasional reversal in high m, n settings). Secondly, M_2 and M_3 show greater RMSEs in general than M_0 and M_1 (which could be ascribed to the former two models exhibiting greater smoothness at d^0), and interestingly enough greater RMSEs than M_4 as well (which is a misspecified model as it returns to 0 at the right end of its support). As expected, increasing both m and n leads to improved performance, by and large. The use of the sigmoidal approximation in the more modest “signal-to-noise” regime ($\sigma = 0.3$) for combinations of larger m and n leads to better results for the smoother models M_2 and M_3 , but not for M_4 , since the shape of the sigmoidal curve conflicts badly with that of the corresponding expected p -value curve. Further, the RMSEs for M_2 generally tend to be worse than M_3 for both $\sigma = 0.1$ and $\sigma = 0.3$, even though M_3 is smoother in the vicinity of d^0 . This can be explained by the fact that for the most part to the right of d^0 , M_3 provides more signal than M_2 , which makes detection easier in the former case. As m and n increase, M_2 improves and its performance comes closer to that of M_3 (see

Table 2: RMSEs for d^0 and τ_0 with $\sigma = 0.1$ and 0.3 , respectively for models M_3 and M_4 and different choices of m and n , using the two proposed methods.

M_3		$\sigma = 0.1$				M_3		$\sigma = 0.3$			
(m, n)		$d^0(1)$	$d^0(2)$	$\tau_0(1)$	$\tau_0(2)$	(m, n)		$d^0(1)$	$d^0(2)$	$\tau_0(1)$	$\tau_0(2)$
(5, 5)		0.091	0.105	0.040	0.032	(5, 5)		0.180	0.161	0.120	0.097
(5, 10)		0.086	0.094	0.027	0.023	(5, 10)		0.191	0.158	0.099	0.067
(10, 10)		0.070	0.081	0.019	0.016	(10, 10)		0.137	0.121	0.067	0.048
(10, 20)		0.078	0.079	0.014	0.011	(10, 20)		0.141	0.122	0.049	0.030
(10, 50)		0.087	0.084	0.009	0.006	(10, 50)		0.147	0.135	0.033	0.019
(20, 50)		0.077	0.074	0.006	0.004	(20, 50)		0.120	0.112	0.021	0.013
(50, 100)		0.072	0.071	0.003	0.002	(50, 100)		0.102	0.099	0.009	0.006

(50,100) setting).

Comparisons between the one-parameter stump and the three-parameter (adaptive) stump: We compared the 1-parameter stump (with 0.5 and 0 as the levels on either side of the unknown split-point) to the three-parameter stump, where both levels as well as the split-point are kept unspecified. The comparison is done for three of the six models at two different σ 's, 0.1 and 0.3, for a set of different (m, n) allocations. Although the results are not shown due to space considerations, the results to a large extent are not radically different, and neither method systematically outperforms the other. For $\sigma = 0.1$ and relatively small m, n , the adaptive stump exhibits slightly better performance for M_1 and M_5 and is more or less comparable to the 1-parameter stump for M_2 . At $\sigma = 0.3$, the advantage of the adaptive stump lessens. Finally, for large (m, n) , both types of stumps behave similarly. Hence, in all subsequent simulations one-parameter stumps are employed.

Assessment of Methods 1 and 2 for estimating d^0 when τ_0 is unknown: We next present a comparison of Methods 1 and 2 from Section 3 for estimating d^0 , when τ_0 is assumed unknown, for two models (due to space considerations): the monotone M_3 , and the tent-shaped M_4 . For each model, we present the RMSEs for d^0 for each of the methods and also those for τ_0 in Table 2.

It is noted that for fixed σ and (m, n) , the RMSEs for d^0 increase from M_3 to M_4 . This phenomenon is also observed when we compare any one of the monotone models M_1 , M_2 or M_3 with either of the tent-shaped ones M_4 or M_5 . Since the stump is a monotone working model, this result is expected in light of the first three models and the lack of monotonicity of the others. Overall, Methods 1 and 2 are comparable, although for smaller $\sigma = 0.1$, Method 1 exhibits smaller RMSEs for d^0 , while for larger $\sigma = 0.3$, Method 2 dominates slightly. It is worth noting that for model M_4 , for small m , the performance deteriorates with increasing n , for large σ . Although counter-intuitive, this phenomenon can be explained by noticing that in this situation the expected p -value curve has a “V”-shape to the right of d^0 and conforms badly with the monotone nature of the fitted stump. For small n , this discrepancy is somewhat masked by the small number of observations (i.e., p -values), but for large n , this non-conformity is clearly exhibited. Regarding τ_0 , Method 2 has a slight edge over Method 1 at both noise levels. Finally, for the challenging setting of M_4 , the performance of both methods is rather inferior for small values of m and $\sigma = 0.3$.

Comparisons among the methods for extreme values of d^0 : We have so far concentrated on the case with $d^0 = 0.5$. We investigate next, settings where d^0 is closer to one of the boundaries; specifically, where the “action” starts fairly quickly at a low covariate level and also where the “action” starts late. To this end, we consider the models \tilde{M}_1 and \tilde{M}_2 , where \tilde{M}_1 is flat at zero till 0.2 and then rises linearly with slope 1 all the way up to 1, while \tilde{M}_2 is

M_4		$\sigma = 0.1$			
(m, n)		$d^0(1)$	$d^0(2)$	$\tau_0(1)$	$\tau_0(2)$
(5, 5)		0.138	0.112	0.067	0.032
(5, 10)		0.109	0.113	0.045	0.026
(10, 10)		0.074	0.102	0.027	0.018
(10, 20)		0.046	0.060	0.018	0.011
(10, 50)		0.034	0.033	0.012	0.007
(20, 50)		0.025	0.028	0.008	0.005
(50, 100)		0.014	0.015	0.003	0.002

M_4		$\sigma = 0.3$			
(m, n)		$d^0(1)$	$d^0(2)$	$\tau_0(1)$	$\tau_0(2)$
(5, 5)		0.243	0.221	0.110	0.096
(5, 10)		0.311	0.294	0.091	0.077
(10, 10)		0.262	0.232	0.081	0.058
(10, 20)		0.286	0.270	0.064	0.047
(10, 50)		0.313	0.317	0.053	0.044
(20, 50)		0.133	0.137	0.040	0.021
(50, 100)		0.050	0.040	0.016	0.006

Table 3: RMSEs in the kink model with $d^0 = 0.2$ (top panels) and 0.8 (bottom panels) for different choices of m and n in both τ_0 known and unknown cases, for $\sigma = 0.1$ and 0.3 respectively.

$d^0 = 0.2$		$\sigma = 0.1$		
(m, n)		d^0	$d^0(1), d^0(2)$	$\tau_0(1), \tau_0(2)$
(5, 5)		0.102	0.333, 0.057	0.327, 0.048
(5, 10)		0.082	0.291, 0.060	0.260, 0.038
(10, 10)		0.081	0.259, 0.054	0.238, 0.027
(10, 20)		0.054	0.143, 0.045	0.127, 0.019
(10, 50)		0.031	0.044, 0.033	0.026, 0.011
(20, 50)		0.027	0.025, 0.027	0.012, 0.008
(50, 100)		0.015	0.015, 0.015	0.005, 0.003

$d^0 = 0.2$		$\sigma = 0.3$		
(m, n)		d^0	$d^0(1), d^0(2)$	$\tau_0(1), \tau_0(2)$
(5, 5)		0.128	0.429, 0.196	0.349, 0.138
(5, 10)		0.120	0.436, 0.181	0.329, 0.111
(10, 10)		0.095	0.396, 0.122	0.317, 0.080
(10, 20)		0.083	0.367, 0.114	0.287, 0.058
(10, 50)		0.080	0.308, 0.112	0.229, 0.036
(20, 50)		0.060	0.210, 0.075	0.161, 0.024
(50, 100)		0.039	0.058, 0.045	0.024, 0.010

flat all the way till 0.8 and then rises linearly with unit slope. These are simple variants of the kink-model M_1 . We report the performances of the 1 parameter stump when τ_0 is known (to be 0) with (the homoscedastic error variance) $\sigma = 0.1$ and 0.3 for both models, and also the performances of Methods 1 and 2 in the τ_0 unknown case for different (m, n) allocations in Table 3. The design-points are chosen similarly to the case $d^0 = 0.5$.

For τ_0 known, the one-parameter stump behaves qualitatively as one would expect. The RMSEs tend to be bigger for the $d^0 = 0.8$ case (for a fixed allocation and noise level), though with increasing (m, n) the RMSEs for both models tend to converge and are similar to the numbers for M_1 in the τ_0 known case, since the models M_1 , \tilde{M}_1 and \tilde{M}_2 look exactly similar in a small neighborhood of the kink, and it is this local behavior that drives the asymptotic MSE as m, n go to infinity. More interesting is the case when τ_0 is unknown. In this case, with the model \tilde{M}_2 , Method 1 generally produces slightly smaller RMSEs for d^0 for both $\sigma = 0.1$ and 0.3 apart from some of the “rich allocation” scenarios where the performances of both methods are very comparable. Method 2, on the other hand, gives better RMSEs for τ_0 . With Model \tilde{M}_1 , we see a markedly different phenomenon. For small (m, n) in the case $\sigma = 0.1$ and $\sigma = 0.3$ (and also for some small m , large n scenarios in the latter case), Method 1 shows very poor performance compared to Method 2 with much larger RMSEs for d^0 . A glance at the RMSEs for τ_0 reveals what is happening. The estimates of τ_0 (needed in Method 1 to compute surrogate p -values) are extremely poor, and this leads to biased estimates for d^0 . The poor performance is, of course, exacerbated for higher values of σ . These results indicate that Method 1 can perform pretty badly for small n ; in that case, the number of covariate values in the flat stretch is small for \tilde{M}_1 and this affects the estimation of τ_0 badly.

A related allocation problem: In our simulation study, the proposed procedures for both known and unknown τ_0 were evaluated for a combination of m and n values. However, in practice one is given a *total budget* of $N \equiv n \times m$ samples that need to be *allocated* to n covariate values and m replicates at each covariate value, respectively. Intuitively, increasing the number of replicates m decreases the “bias”, whereas increasing the number of values n

$d^0 = 0.8$		$\sigma = 0.1$			$d^0 = 0.8$		$\sigma = 0.3$		
(m, n)		d^0	$d^0(1), d^0(2)$	$\tau_0(1), \tau_0(2)$	(m, n)		d^0	$d^0(1), d^0(2)$	$\tau_0(1), \tau_0(2)$
(5, 5)		0.213	0.128, 0.174	0.028, 0.028	(5, 5)		0.178	0.114, 0.146	0.080, 0.078
(5, 10)		0.119	0.088, 0.118	0.023, 0.019	(5, 10)		0.115	0.099, 0.109	0.060, 0.051
(10, 10)		0.126	0.095, 0.114	0.016, 0.013	(10, 10)		0.120	0.092, 0.111	0.044, 0.037
(10, 20)		0.060	0.050, 0.057	0.011, 0.008	(10, 20)		0.087	0.088, 0.092	0.033, 0.025
(10, 50)		0.033	0.033, 0.035	0.008, 0.005	(10, 50)		0.084	0.097, 0.091	0.024, 0.016
(20, 50)		0.026	0.025, 0.028	0.005, 0.004	(20, 50)		0.060	0.068, 0.064	0.017, 0.011
(50, 100)		0.016	0.015, 0.015	0.002, 0.002	(50, 100)		0.038	0.042, 0.039	0.008, 0.005

Table 4: Optimal allocation (m, n) pairs for a fixed total budget $N = m \times n$

M1	Method 1	$\sigma = 0.1$	$\sigma = 0.3$	Method 2	$\sigma = 0.1$	$\sigma = 0.3$
	$N = 100$	(6,17)	(33,3)	$N = 100$	(4,25)	(33, 3)
	$N = 200$	(7,29)	(40,5)	$N = 200$	(6,33)	(15,13)
M5	Method 1	$\sigma = 0.1$	$\sigma = 0.3$	Method 2	$\sigma = 0.1$	$\sigma = 0.3$
	$N = 100$	(8,12)	(33,3)	$N = 100$	(5,20)	(33,3)
	$N = 200$	(7,29)	(67,3)	$N = 200$	(6,33)	(15,13)

of the covariate, decreases the variance of the estimators. The optimal allocation occurs when the two terms are balanced, usually at a moderate value of n and m (which depends on the value of σ and the regression function). Thus, for a fixed N , one expects that the RMSEs exhibit a “U-shape” as a function of m ; further, for larger σ the optimal allocation would occur at a larger value of m .

We investigate this allocation problem through a simulation, but due to space considerations we present the optimal allocations for models M1 and M5 for both Methods 1 and 2. The setting under consideration is $d^0 = 0.5$, $N = 100$ and 200 and $\sigma = 0.1$ and 0.3 . All possible combinations of m and n that approximately satisfy the total budget were considered. The optimal allocations are shown in Table 4.

It can be seen that both methods for small σ favor lots of covariate values and few replicates, while the situation is reversed for high σ . Further, qualitatively similar results, in accordance with our observation above, are obtained for the other three models examined (M_2 , M_3 and M_4). Nevertheless, a few anomalies are present; specifically, as we are sampling from the discrete uniform design on $[0, 1]$, and $d^0 = 0.5$, sometimes the optimal allocation occurs at the rather extreme value $n = 3$. This is due to the fact that in that case, the covariate values are placed at 0.25, 0.5 and 0.75, and when m is large, the fitted break point \hat{d}_n is usually (for many replicates) 0.5, the true parameter value. Whenever this is the case, the estimation error is exactly zero, making the observed RMSEs small. With the same budget, a larger n (say $n = 5$) can also lead to 0.5 as a covariate value, but the value of m decreases in the process (thereby increasing the bias) and there are more options for the fitted break point to differ from 0.5, leading to larger RMSEs.

Some practical recommendations: Based on our extensive simulation study (including results not shown here due to space considerations), the following practical recommendations are in order. Overall, it is better for one to invest in an increased number of covariate values (n), rather than replicates (m). In the case of a known τ , the simple stump model performs well overall, while the more complicated adaptive stump model offers only marginal improvements. In the case where the threshold d^0 is closer to the boundaries, investment in n proves fairly important. For unknown level τ_0 , none of the proposed methods dominates the other, the result depending on both the noise level and the model under consideration. However, for estimation of the level τ_0 , Method 2 exhibits a clear advantage over its competitor. Some

Table 5: RMSEs for the five procedures for different choices of m and n when $\sigma = 0.3$ and the actual model is M_1 (left table) and M_2 (right table).

(m, n)	P_1	P_2	P_3	P_4	P_5	P_1	P_2	P_3	P_4	P_5
(5, 5)	0.163	0.207	0.339	0.255	0.299	0.204	0.241	0.420	0.291	0.298
(5, 10)	0.134	0.176	0.304	0.307	0.344	0.201	0.227	0.390	0.346	0.360
(10, 10)	0.119	0.120	0.227	0.228	0.328	0.168	0.194	0.334	0.302	0.360
(10, 20)	0.092	0.079	0.191	0.265	0.295	0.177	0.163	0.303	0.329	0.354
(10, 50)	0.085	0.042	0.179	0.310	0.247	0.193	0.150	0.294	0.369	0.332
(20, 50)	0.060	0.030	0.128	0.212	0.176	0.162	0.147	0.245	0.305	0.274
(50, 100)	0.038	0.013	0.080	0.142	0.114	0.132	0.145	0.197	0.254	0.211

Table 6: RMSEs for the five procedures for different choices of m and n when $\sigma = 0.3$ and the actual model is M_3 (left table) and M_5 (right table).

(m, n)	P_1	P_2	P_3	P_4	P_5	P_1	P_2	P_3	P_4	P_5
(5, 5)	0.173	0.197	0.342	0.245	0.314	0.181	0.232	0.376	0.277	0.286
(5, 10)	0.140	0.159	0.306	0.297	0.351	0.153	0.239	0.370	0.370	0.287
(10, 10)	0.126	0.116	0.247	0.228	0.319	0.117	0.203	0.240	0.336	0.282
(10, 20)	0.117	0.084	0.216	0.256	0.282	0.093	0.191	0.198	0.411	0.266
(10, 50)	0.129	0.068	0.203	0.302	0.248	0.084	0.168	0.178	0.465	0.241
(20, 50)	0.110	0.064	0.170	0.213	0.194	0.060	0.148	0.127	0.440	0.175
(50, 100)	0.098	0.060	0.138	0.164	0.151	0.038	0.139	0.080	0.402	0.113

hybrid possibilities are discussed in the concluding remarks section.

5.2 Comparison with other procedures

Next, we compare the proposed 1-parameter stump method to some competing procedures developed in the pharmacological dose-response setting to identify the MED. Most of the methods developed in dose-response setting context are based on hypothesis testing procedures. For example, Williams (1971) developed a method to identify the lowest dose at which there is “activity” in toxicity studies using a closed testing procedure based on isotonic regression for a monotone dose-response relationship. Hsu and Berger (1999) developed a step-wise confidence set approach to estimate and make inference on the MED. A nonparametric method based on the Mann-Whitney statistic incorporating the step-down procedure is investigated in Chen (1999), while Tamhane and Logan (2002) use multiple testing procedures for the task at hand. We compare our method with that of Williams (1971), of Hsu and Berger (1999) and of Chen (1999), referred henceforth as P_3 , P_4 and P_5 , respectively.

We fit the stump ξ_d with levels 1/2 and 0 on either side of the threshold d to the observed p -values (with $\tau_0 \equiv 0$ assumed known) and compare the performance of our approach P_1 with that of P_3 , P_4 and P_5 . A natural parametric procedure to estimate d^0 might be to fit a kink-type (hockey stick) model like M_1 to the observed responses and estimate d (the threshold parameter) and the slope of the linear segment by the least squares method. We also implement this method and call it P_2 . Obviously when the true underlying regression function μ is not a kink-model this method might not be consistent, but given a finite sample it is often a good first approximation. Whereas, when μ is a kink-function, e.g., when we assume the true model to be M_1 , this approach should clearly outperform the other procedures. Indeed, Table 5 shows that P_2 is very competitive for model M_1 ; still our approach P_1 performs better for small sample sizes, e.g., (5, 5) and (5, 10). For the model M_2 , a slight departure from the model M_1 , P_1 mostly dominates P_2 , and all the other procedures. Note that as P_3 , P_4 and P_5 are procedures that are based on testing hypotheses, we need to specify a level (α), and in the simulations reported in the paper we have set $\alpha = 0.05$. The choice of the $\alpha = 0.05$ is purely based on classical hypothesis testing considerations; a proper choice

of the tuning parameter is not available. Changing α will change the RMSEs of P_3 – P_5 , and it is not quite clear how that will affect the estimation procedure and the RMSEs. Also, to implement P_3 – P_5 , we computed the cut-off values necessary to carry out the hypothesis tests using simulation, as such tables are not available for the different choices of m and n considered in this paper.

Table 6 shows the performance of the methods when the true models are M_3 , the infinitely differentiable regression function, and M_5 . Clearly P_1 dominates all the other methods when the data is generated according to M_5 ; notice that P_3 and P_4 work with the underlying assumption that μ is nondecreasing, a condition which is violated in M_5 , and this explains the poor performance of these methods. Also, P_2 is biased in this scenario, and thus the RMSEs do not converge to 0 for large m and n . When M_3 is the true model, P_2 performs surprisingly well; this can be explained by looking at Figure 4, the threshold value for M_3 is approximated very well by that of the kink-model. This is an artifact of the particular choice of the infinitely differentiable function, and the RMSEs can in general be very different if the μ is not well approximated by a kink function. Overall, P_1 is very competitive, and the simplicity of our approach coupled with its adaptivity to different types of mis-specifications, makes it a very attractive choice. We also note that the fitting of ξ_d does not require any tuning parameter, an obvious advantage over the testing based procedures. Indeed, one of the novelties of our approach lies in the fact that we treat the estimation of d^0 purely as an estimation problem and not a result of a series of hypotheses tests, thereby avoiding the need to specify α .

5.3 Data Applications

In this section, we apply the proposed procedure to the two motivating applications. Note that the first one corresponds to a rich allocation scheme in terms of (m, n) , while the second application to a sparse one, thus showing the range of applicability of the procedure. Further, for the first application we discuss a subsampling mechanism that allows us to calculate confidence intervals for \hat{d}_0 , given the large number of doses available. This is not repeated for the gene expression data due to the paucity of time points, but for richer time course experiments it would become applicable.

Queueing System: We consider a complex system comprising multiple classes of customers waiting at infinity capacity queues and a set of processing resources modulated by an external stochastic process. The system employs a resource allocation (scheduling) policy that decides at every time slot which customer class to serve, given the state of the modulating rate process and the backlog of the various queues. In Bambos and Michailidis (2004), a low complexity policy was introduced and its maximum throughput properties established. This canonical system captures the essential features of data/voice transmissions in a wireless network, in multi-product manufacturing systems, and in call centers (for more details see Bambos and Michailidis (2004)). As discussed in the introductory section, an important quantity of interest to the system’s operator is the average delay of jobs (over all classes), which constitutes a key performance metric of the quality of service offered by the system. The average delay of the jobs in a two-class system as a function of its loading under the optimal policy, for a small set of loadings is shown in the left panel of Figure 1. These responses were obtained through simulation, since for such complex systems analytic calculations are intractable.

We next employ the developed methodology for estimating both the loading d^0 and the unknown level τ_0 . Ten replicates of the response (average delay) were obtained based on 5,000 events per class by simulating the system under consideration and after accounting for a burn-in period of 2,000 per class in order to ensure that it reached its stationary regime. The means per loading, \bar{Y}_i s, are shown in Figure 5.

We applied both procedures discussed in Section 3 with an unknown value for τ_0 , assum-

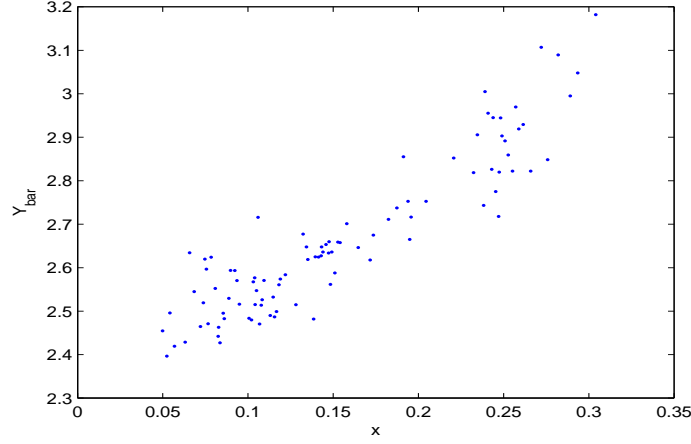


Figure 5: Plot of the average responses \bar{Y}_i .

ing heteroscedastic errors. The first method (Method 1) gives an estimate of τ_0 , denoted by $\hat{\tau}_0(1) = 2.61$ with corresponding $\hat{d}^0(1) = 0.151$, while the second method (Method 2) gives $\hat{\tau}_0(2) = 2.52$ with corresponding $\hat{d}^0(2) = 0.117$. It can be seen that there is fairly strong agreement between the two estimates. From the system's operator point of view the average delay of jobs exhibits a markedly increasing trend beyond a loading of 15%. For the first method, a plot of the P-values (left panel) and the criterion function (right panel) that is minimized in Theorem 3.2 are given in Figure 6.

A second analysis, assuming homoscedastic errors, produces fairly comparable results: $\hat{d}^0(1) = 0.151$ and $\hat{\tau}_0(1) = 2.59$ and $\hat{d}^0(2) = 0.128$ and $\hat{\tau}_0(2) = 2.52$, respectively for the two methods. Another question of interest to the system's operation is what level of uncertainty, as reflected

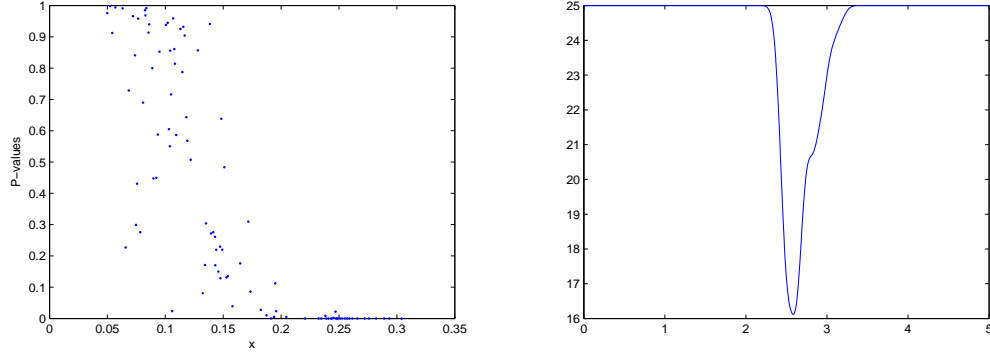


Figure 6: Plots of the estimated p -values using Method 1 (left panel) and the plot of the criterion function whose minimizer provides the estimate of τ_0 (right panel).

through confidence intervals, can be assigned to these estimates. Obviously, our results establish consistency of the threshold \hat{d}^0 and level $\hat{\tau}_0$ estimates, but no characterization of their asymptotic distribution has been provided, which would have resolved this issue. Nevertheless, we outline a subsampling based procedure for partially addressing the construction of confidence intervals, provide some theoretical justification in the Appendix and finally discuss some open issues (see Section 6).

The steps in the employed procedure are:

1. Sample m_n vectors out of the 100 $\{X_i, \bar{Y}_i\}$ pairs without replacement.
2. For the collected sub-sample, compute estimates of d^0 via the two methods, denoted by $\hat{d}^{0,*}(1)$ and $\hat{d}^{0,*}(2)$ respectively. Two versions of $\hat{d}^{0,*}(1)$ are calculated: the first, denoted by $\hat{d}^{0,*}(1, a)$ is calculated using the estimate of τ_0 based on the full sample (in other words, taking $\hat{\tau}_0(1)$ as the “truth”), while the second denoted by $\hat{d}^{0,*}(1, b)$ is computed after re-estimating τ_0 from the obtained subsample.
3. Calculate the following statistics: $t_{1n}^* \equiv m_n^{1/3} (\hat{d}^{0,*}(1, a) - \hat{d}^0(1))$, $t_{2n}^* \equiv m_n^{1/3} (\hat{d}^{0,*}(1, b) - \hat{d}^0(1))$ and $t_{3n}^* \equiv m_n^{1/3} (\hat{d}^{0,*}(2) - \hat{d}^0(2))$.
4. Repeat the above 3 steps a large number of times (say B), storing the three statistics in the preceding step for each iteration, and obtain the empirical distributions for each of these statistics based on the B iterates, say $\{F_{j,n}^*\}_{j=1}^3$.
5. Calculate $q_{.025,j,n}^*$ and $q_{.975,j,n}^*$, the 2.5-th and 97.5-th percentiles of $F_{j,n}^*$ respectively, for $j = 1, 2, 3$.
6. Prescribe $[\hat{d}^0(1) - n^{-1/3} q_{.975,j,n}^*, \hat{d}^0(1) - n^{-1/3} q_{.025,j,n}^*]$, for $j = 1, 2$, and $[\hat{d}^0(2) - n^{-1/3} q_{.975,3,n}^*, \hat{d}^0(2) - n^{-1/3} q_{.025,3,n}^*]$ as approximate 95% confidence intervals for d^0 .

Using $m_n = 50$ and assuming heteroscedastic errors the 95% confidence intervals for d^0 when $j = 1, 2, 3$, are (.133, .179), (.135, .158) and (.099, .125) respectively. Very similar results are obtained for $m_n = 75$, while for smaller values of m_n (e.g. 10 or 25) the resulting confidence intervals become exceedingly wide. Under the assumption of homoscedasticity the obtained 95% confidence intervals are (0.133, 0.178), (0.151, 0.176) and (0.120, 0.139), respectively. It should be noted that the third confidence interval does not overlap with the second one, something that can be attributed to their fairly narrow respective lengths. From the system's operator perspective, it can be seen that running the system at loadings below 12% of the total capacity produces very small delays on the average for its customers, while at loadings larger than 18%, customers should expect to experience increasing delays.

Time course analysis of the Transglutaminase 2 gene: This example deals with a time course experiment that studies the effects of cells treated with TGF- β , a key cytokine implicated in a number of disease processes including cancer, on the epithelial-mesenchymal transition, a phenotypic conversion that enables cancer cells to attain their migratory and invasive capacities (see Keshamouni et al. (2009)). The data correspond to the expression levels of the Transglutaminase 2 gene measured in triplicate at 0, 0.5, 1, 2, 4, 8, 16, 24 and 72 hours. This gene has been implicated in this mechanism through an elevated level at 72 hours, compared to the baseline at 0 hours (see Keshamouni et al. (2006)). The questions of interest are (a) at what time point its expression level rises from its baseline, since this provides a timeline for the activation of this transition mechanism in cell-lines; (b) at what time point its expression level reaches saturation, which indicates that transition to malignancy is essentially complete.

The average expression level of the 3 replicates is shown in Figure 7 and it can be seen that it rises from its initial baseline level and subsequently flattens out. It is therefore reasonable to postulate a model that is constant (τ_{\min}) till the first transition point d_{\min} , then increases monotonically until it reaches the second transition point d_{\max} beyond which it remains constant at τ_{\max} . Without loss of generality, for the analysis, the nine design points were taken as equispaced in the $[0, 1]$ interval, since our interest focuses on identifying the stages where the changes occur. The procedure described in Section 4 was employed to estimate the parameters of interest with the p -values being computed under the assumption of

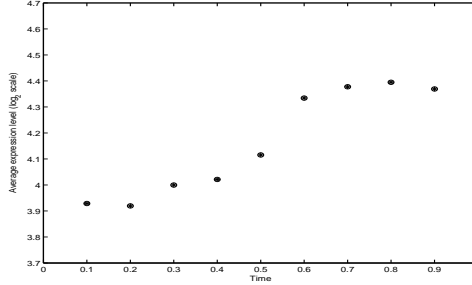


Figure 7: Average expression of the Transglutaminase 2 gene over time.

heteroscedasticity of the error (as evident from the right panel of Figure 1). For the baseline, Method 1 gives an estimate $\hat{\tau}_{\min}(1) = 3.974$ with corresponding $\hat{d}_{\min}(1) = 0.4$, while Method 2 gives $\hat{\tau}_{\min}(2) = 3.924$ with corresponding $\hat{d}_{\min}(2) = 0.2$. It can be seen that the two methods basically agree on the value of τ_{\min} , but the second one identifies the change-point somewhat earlier than the first method. For the maximum level, Method 1 gives an estimate $\hat{\tau}_{\max}(1) = 4.367$ with corresponding $\hat{d}_{\max}(1) = 0.6$, while Method 2 gives $\hat{\tau}_{\max}(2) = 4.369$ with corresponding $\hat{d}_{\max}(2) = 0.6$, thus exhibiting perfect agreement. To estimate the minimum (baseline) value using Method 1 we minimized (9) over the restricted interval $[\min(\bar{Y}_{i.}), 4]$ and for the maximum over the restricted interval $(4, \max(\bar{Y}_{i.})]$ respectively, as advocated in Section 4. The choice of the intervals do not matter so long they are disjoint and are contain the true minimum and maximum. We believe that the 0.4 time point is a more accurate estimate of d_{\min} , since the proteins encoded by the Transglutaminase 2 gene exhibit increased levels at 8 hours (the 0.5 design point) and nothing significant prior to that time, as obtained from a separate experimental platform described in Keshamouni et al. (2009).

6 Concluding Discussion

In this paper, we address the problem of identifying the threshold parameter at which a regression function diverges from its (possibly unknown) baseline value employing a p -value framework, under a controlled sampling setting. The p -values exhibit a natural dichotomy in behavior on different sides of the threshold, which is crucially used in our methodology. Our approach is computationally simple, nonparametric in nature and adaptive in the sense that it does not need any specification of the local behavior of the regression function around the threshold value. The procedure can also be used to prescribe estimates for the baseline value of the regression function. We establish asymptotic properties of the proposed estimators, extend them to the case of multiple change-points, study their finite sample behavior through an extensive simulation study, compare them for the MED problem to competing approaches and apply them to two real applications. The numerical results indicate that the procedure performs well in both high and relatively low signal-to-noise settings.

We conclude with a discussion of a number of issues, some of which will be the focus of future work. In our setting, we dealt with the case of a balanced design with a fixed number of replicates m for every dose level X_i . The case of varying number of replicates m_i can be handled analogously, although some care needs to be exercised in the technical arguments. Under the assumption that the minimum of the m_i 's goes to infinity, all consistency results in this paper continue to hold. A natural extension of the problem would be to investigate the performance of our method when we have a random design of points with no replicates, i.e., $m_i = 1$. Such problems arise in diverse contexts, e.g., in the astronomy application discussed in the Introduction, where the goal is to estimate the “tidal” radius. Note that in this setup, (approximate) p -values can be computed by binning the covariate space and averaging the

responses in each bin.

The problem of constructing confidence intervals for the threshold d^0 is of considerable interest in many applications. *Ideally*, one would like to consider the asymptotic distribution of $\hat{d}_{m,n}$, as both m, n increase to infinity and use the quantiles of the resulting distribution to calibrate confidence intervals. However, as discussed in the previous section, this is a hard problem and its solution, presently unknown, is outside the scope of this paper. This constitutes a subject of future research.

In the case of an unknown threshold τ_0 , our numerical results show a superior performance of Method 2 over its competitor, especially in small (m, n) scenarios. The poorer performance of Method 1 in these situations was tracked down to the inaccurate estimation of τ_0 in the simulation section. Note that consistency of the estimator is formally established for Method 1. It would therefore be interesting to explore a hybrid procedure and its properties, where τ_0 is estimated using Method 2 and the resulting estimate is used as a plug-in in Method 1.

Although we develop our method in a simple univariate regression setup, our approach, can be generalized to identify the “baseline” region in multi-dimensional covariate spaces. Further, for regression models in higher dimensions, the problem of estimating regions in the covariate space where the regression function stays below a pre-specified threshold can also be handled by our approach. Such problems, known as *level-sets estimation*, have been extensively studied in the statistics and engineering literature (see for example, Singh et al. (2009), Willet and Nowak (2007)). Indeed, Section 2.2 of the current paper may be viewed as a level-set estimation problem, albeit in a simpler setting, but under minimal assumptions on the behavior of the regression function at the boundary.

In conclusion we recall that the problem treated in this paper could also have been treated by direct estimation of the underlying regression function μ . However, as briefly mentioned in the Introduction, straightforward intuitive estimates of the form $\hat{d} \equiv \inf\{x : \hat{\mu}(x) > 0\}$ underestimates d^0 ; potential fixes lead to solutions for which very little is known in terms of asymptotic properties. Nevertheless, it would be fruitful to explore that line of research.

Acknowledgements

We would like to thank Harsh Jain for bringing to our attention a threshold estimation problem that eventually led to the formulation and development of this framework. The work of the authors were partially supported by NSF and NIH grants: DMS-09-06597 (BS), DMS-07-05288 (MB) and 1RC1CA145444-0110 (GM).

7 Appendix

We start with establishing an auxiliary result used in subsequent developments.

Theorem 7.1 *Let $\{\mathbb{M}_n^\tau : \tau \in \mathcal{T}\}_{n=1}^\infty$ be a family of (real-valued) stochastic processes indexed by $h \in \mathcal{H}$ and let $\{M^\tau : \tau \in \mathcal{T}\}$ be a family of deterministic functions defined on \mathcal{H} , such that each M^τ is minimized at a unique point $h(\tau) \in \mathcal{H}$. Here \mathcal{H} is a metric space and denote the metric on \mathcal{H} by d . Let \hat{h}_n^τ be a minimizer of \mathbb{M}_n^τ . Assume further that:*

- (a) $\sup_{\tau \in \mathcal{T}} \sup_{h \in \mathcal{H}} |\mathbb{M}_n^\tau(h) - M^\tau(h)| \xrightarrow{P} 0$ and
- (b) For every $\eta > 0$, $c(\eta) \equiv \inf_{\tau} \inf_{h \notin B_\eta(h(\tau))} \{M^\tau(h) - M^\tau(h(\tau))\} > 0$, where $B_\eta(h)$ denotes the open ball of radius η around h .

Then, (i) $\sup_{\tau} d(\hat{h}_n^\tau, h^\tau)$ converges in probability to 0. Furthermore, if \mathcal{T} is a metric space and h^τ is continuous in τ , then (ii) $\hat{h}_n^{\tau_n} \xrightarrow{P} h^{\tau_0}$, provided τ_n converges to τ_0 . In particular, if

the \mathbb{M}_n^τ 's themselves are deterministic functions, the conclusions of the theorem hold with the convergences in probability in (i) and (ii) replaced by usual non-stochastic convergence.

Proof: We provide the proof in the case that \mathcal{H} is a sub-interval of the real line, the case that is relevant for our applications. However, there is no essential difference in generalizing to the metric space case. Euclidean distances simply need to be replaced by the metric space distance and open intervals by open balls.

Given $\eta > 0$, we need to deal with $P^* \{\sup_{\tau \in \mathcal{T}} |\hat{h}_n^\tau - h(\tau)| > \eta\}$. We deal with outer probabilities to avoid measurability difficulties. The event $A_{n,\eta} \equiv \{\sup_{\tau \in \mathcal{T}} |\hat{h}_n^\tau - h(\tau)| > \eta\}$ implies that for some τ , $\hat{h}_n^\tau \notin (h(\tau) - \eta, h(\tau) + \eta)$ and therefore

$$M^\tau(\hat{h}_n^\tau) - M^\tau(h(\tau)) \geq \inf_{h \notin (h(\tau) - \eta, h(\tau) + \eta)} \{M^\tau(h) - M^\tau(h(\tau))\}.$$

This is equivalent to

$$\begin{aligned} M^\tau(\hat{h}_n^\tau) - M^\tau(h(\tau)) - \mathbb{M}_n^\tau(\hat{h}_n^\tau) + \mathbb{M}_n^\tau(h(\tau)) &\geq \\ \inf_{h \notin (h(\tau) - \eta, h(\tau) + \eta)} \{M^\tau(h) - M^\tau(h(\tau))\} + \mathbb{M}_n^\tau(h(\tau)) - \mathbb{M}_n^\tau(\hat{h}_n^\tau). \end{aligned}$$

Now, $\mathbb{M}_n^\tau(h(\tau)) - \mathbb{M}_n^\tau(\hat{h}_n^\tau) \geq 0$ and the left side of the above display is dominated by

$$2 \|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \equiv 2 \sup_{h \in \mathcal{H}} |\mathbb{M}_n^\tau(h) - M^\tau(h)|,$$

implying that:

$$2 \|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq \inf_{h \notin (h(\tau) - \eta, h(\tau) + \eta)} \{M^\tau(h) - M^\tau(h(\tau))\},$$

which, in turn, implies that:

$$2 \sup_{\tau \in \mathcal{T}} \|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq \inf_{\tau \in \mathcal{T}} \inf_{h \notin (h(\tau) - \eta, h(\tau) + \eta)} \{M^\tau(h) - M^\tau(h(\tau))\} \equiv c(\eta),$$

by definition. Hence

$$A_{n,\eta} \subset \left\{ \sup_{\tau \in \mathcal{T}} \|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq c(\eta)/2 \right\}.$$

By assumptions (a) and (b), $P^* \{\sup_{\tau \in \mathcal{T}} \|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq c(\eta)/2\}$ goes to 0 and therefore so does $P^*(A_{n,\eta})$. \square

Remarks: We will call the sequence of steps involved in deducing the inclusion:

$$\left\{ \sup_{\tau \in \mathcal{T}} |\hat{h}_n^\tau - h(\tau)| > \eta \right\} \subset \left\{ \sup_{\tau \in \mathcal{T}} \|\mathbb{M}_n^\tau - M^\tau\|_{\mathcal{H}} \geq c(\eta)/2 \right\},$$

as *generic steps*. Very similar steps will be required time and again at places in the proofs of the theorems to follow. We will not elaborate those arguments, but refer back to the *generic steps* in such cases.

Proof of Theorem 2.1: We prove Part (b) of the theorem since Part (a) follows by an (easier) adaptation of the arguments needed for Part (b). Recall that in Part (b), we find the best fitting stump to the observed p -values $Z_{im}, i = 1, 2, \dots, n$. Letting $\xi_\theta(x) \equiv \alpha 1(x \leq d) + \beta 1(x > d)$ for $\theta = (\alpha, \beta, d)$, we minimize

$$\mathbb{M}_{m,n}(\theta) = \sum_{i=1}^n \{Z_{im} - \xi_\theta(X_i)\}^2 = \sum_{i: X_i \leq d} (Z_{im} - \alpha)^2 + \sum_{i: X_i > d} (Z_{im} - \beta)^2 \quad (11)$$

over $\theta = (\alpha, \beta, d) \in [0, 1]^3$. Letting $\hat{\theta}_{m,n} = (\hat{\alpha}_{m,n}, \hat{\beta}_{m,n}, \hat{d}_{m,n}) \equiv \arg \min_{\theta \in [0, 1]^3} \mathbb{M}_{m,n}(\theta)$, we see that $\hat{\theta}_{m,n}$ is a natural estimator of $\theta_0 \equiv (0.5, 0, d^0)$. Let $\|\cdot\|$ denote the l_∞ -metric in \mathbb{R}^3 , i.e., $\|(a, b, d)\| \equiv \max\{|a|, |b|, |d|\}$. Note that as m changes, the distribution of Z_{im} changes, and so we effectively have a triangular array of i.i.d. random variables $\{(X_i, Z_{im})\}_{i=1}^n \sim P_m$. It suffices to show that $\hat{\theta}_{m,n} \xrightarrow{P} \theta_0$ as $m, n \rightarrow \infty$, i.e., given $\epsilon, \xi > 0$, there exists $K \in \mathbb{N}$, such that for all $m, n \geq K$, $P_m\{\|\hat{\theta}_{m,n} - \theta_0\| > \epsilon\} < \xi$.

Using empirical process notation, note that $\mathbb{M}_{m,n}(\theta) \equiv \mathbb{P}_{n,m}\{Z_{1m} - \xi_\theta(X_1)\}^2$ and define $M_m(\theta) \equiv P_m\{Z_{1m} - \xi_\theta(X_1)\}^2$ where $M_m(\theta)$ can be simplified as

$$M_m(\theta) = \int_0^d \{\nu_m(x) - \alpha\}^2 p_X(x) dx + \int_d^1 \{\nu_m(x) - \beta\}^2 p_X(x) dx + c_m, \quad (12)$$

with $c_m = \int_0^1 \sigma_m^2(x) p_X(x) dx$, where $\sigma_m^2(x) = \text{Var}(Z_{im} | X_i = x)$. Note that $\sigma_m^2(x) \rightarrow (1/12)1(x \leq d^0) \equiv \sigma^2(x)$ as $m \rightarrow \infty$. Let $M(\theta)$ be the same expression for $M_m(\theta)$ in (12) with $\nu_m(x)$ replaced by $\nu(x) = (1/2)1(x \leq d^0)$ and c_m replaced by $c = \int_0^1 \sigma^2(x) p_X(x) dx$, e.g., $M(\theta) = \int_0^d \{\nu(x) - \alpha\}^2 p_X(x) dx + \int_d^1 \{\nu(x) - \beta\}^2 p_X(x) dx + c$. Observe that $M(\theta) \geq c$ for all θ , and $M(\theta_0) = c$. Also it is easy to observe that θ_0 is the unique minimizer of $M(\theta)$. Note that $M_m(\theta)$ can be expanded as

$$\begin{aligned} M_m(\theta) &= \int_0^1 \nu_m^2(x) p_X(x) dx - 2\alpha \int_0^d \nu_m(x) p_X(x) dx \\ &\quad - 2\beta \int_d^1 \nu_m(x) p_X(x) dx + \alpha^2 \int_0^d p_X(x) dx + \beta^2 \int_d^1 p_X(x) dx + c_m. \end{aligned}$$

Using a similar expansion for $M(\theta)$, the difference $|M_m(\theta) - M(\theta)|$, can be bounded as

$$\begin{aligned} &2 \int_0^1 |\nu_m^2(x) - \nu^2(x)| p_X(x) dx + 2\alpha \int_0^d |\nu_m(x) - \nu(x)| p_X(x) dx \\ &- 2\beta \int_d^1 |\nu_m(x) - \nu(x)| p_X(x) dx + |c_m - c| \\ &\leq 8 \int_0^1 |\nu_m(x) - \nu(x)| p_X(x) dx + |c_m - c| \rightarrow 0 \end{aligned}$$

uniformly in $\theta \in [0, 1]^3$, e.g., $\|M_m - M\|_\infty \equiv \sup_{\theta \in [0, 1]^3} |M_m(\theta) - M(\theta)| \rightarrow 0$. By Theorem 7.1, $\theta_m = (\alpha_m, \beta_m, d_m) = \arg \min_{\theta \in [0, 1]^3} M_m(\theta) \rightarrow \arg \min_{\theta \in [0, 1]^3} M(\theta) = \theta_0$ as $m \rightarrow \infty$. Notice now that it is enough to show that for any $\epsilon > 0$, for some M_0 (possibly depending on ϵ),

$$\sup_{m \geq M_0} P_m\{\sup_{n \geq k} \|\hat{\theta}_{m,n} - \theta_m\| > \epsilon\} \rightarrow 0, \text{ as } k \rightarrow \infty. \quad (13)$$

To see this, take any $\xi > 0$. Then, by (13) and the fact that $\theta_m \rightarrow \theta_0$, there exists $K \in \mathbb{N}$, $K > M_0$ such that for all $k \geq K$,

$$\sup_{m \geq M_0} P\{\sup_{n \geq k} \|\hat{\theta}_{m,n} - \theta_m\| > \epsilon/2\} \leq \xi, \text{ and } \|\theta_k - \theta_0\| < \epsilon/2,$$

which implies for $n, m \geq K$,

$$P\{\|\hat{\theta}_{m,n} - \theta_0\| > \epsilon\} \leq P\{\|\hat{\theta}_{m,n} - \theta_m\| > \epsilon/2\} \leq P\{\sup_{n \geq K} \|\hat{\theta}_{m,n} - \theta_m\| > \epsilon/2\} \leq \xi,$$

thereby completing the argument.

To show that (13) holds, consider the class of functions $\mathcal{F} \equiv \{f_\theta(x, z) \equiv (z - \alpha)^2 1(x \leq d) + (z - \beta)^2 1(x > d) | \theta = (\alpha, \beta, d) \in [0, 1]^3\}$ with the envelope $F(x, z) = 1$. Note that \mathcal{F} is formed by combining three bounded VC classes of functions: $\{(z - \alpha)^2 : 0 \leq \alpha \leq 1\}$, $\{(z - \beta)^2 : 0 \leq \beta \leq 1\}$ and $\{1(x \leq d) : 0 \leq d \leq 1\}$ through finitely many operations involving addition and multiplication and therefore satisfies the entropy condition in the third display on page 168 of van der Vaart and Wellner (1996). It follows that \mathcal{F} satisfies the conditions of Theorem 2.8.1 of van der Vaart and Wellner (1996) and is therefore uniformly Glivenko-Cantelli for the class of probability measures $\{P_m\}$, i.e.,

$$\sup_{m \geq 1} P_m \left\{ \sup_{n \geq k} \|\mathbb{M}_{m,n} - M_m\|_\infty > \epsilon \right\} \rightarrow 0 \quad (14)$$

for every $\epsilon > 0$ as $k \rightarrow \infty$, where $\|\cdot\|_\infty$ denotes the supremum (uniform) norm over the function class.

Fix $\epsilon > 0$ and consider the event $\{\|\hat{\theta}_{m,n} - \theta_m\| > \epsilon\}$. Since θ_m minimizes M_m and $\hat{\theta}_{m,n}$ minimizes $\mathbb{M}_{m,n}$, by arguments analogous to the *generic steps* in the proof of Theorem 7.1, we have:

$$\|\hat{\theta}_{m,n} - \theta_m\| > \epsilon \Rightarrow \|\mathbb{M}_{m,n} - M_m\|_\infty \geq \eta_m(\epsilon)/2, \quad (15)$$

where

$$\eta_m(\epsilon) = \inf_{\theta \in [\theta_m - \epsilon \mathbf{1}, \theta_m + \epsilon \mathbf{1}]^c} \{M_m(\theta) - M_m(\theta_m)\}.$$

and $\mathbf{1} = (1, 1, 1)'$.

Claim: There exists $\eta > 0$ and an integer M_0 such that $\eta_m(\epsilon) \geq \eta > 0$ for all $m \geq M_0$. We assume the claim for the time being, which yields,

$$P_m \left\{ \sup_{n \geq k} \|\hat{\theta}_{m,n} - \theta_m\| > \epsilon \right\} \leq P_m \left\{ \sup_{n \geq k} \|\mathbb{M}_{m,n} - M_m\|_\infty > \eta/2 \right\}, \forall m \geq M_0$$

and thus by (14), $\sup_{m \geq M_0} P \left\{ \sup_{n \geq k} \|\hat{\theta}_{m,n} - \theta_m\| > \epsilon \right\} \rightarrow 0$ as $k \rightarrow \infty$. This completes the proof of the theorem. \square

Proof of the Claim: Let us bound $M_m(\theta) - M_m(\theta_m)$ below as

$$\begin{aligned} M_m(\theta) - M_m(\theta_m) &= (M_m - M)(\theta) - (M_m - M)(\theta_m) + \{M(\theta) - M(\theta_m)\} \\ &\geq -2\|M_m - M\|_\infty + \{M(\theta) - M(\theta_m)\} \end{aligned}$$

As $\|M_m - M\|_\infty \rightarrow 0$ as $m \rightarrow \infty$, it is enough to show that there exists $\eta > 0$ such that for all sufficiently large m , $\inf_{\theta \in [\theta_m - \epsilon \mathbf{1}, \theta_m + \epsilon \mathbf{1}]^c} \{M(\theta) - M(\theta_m)\} > \eta$. Note that the main difficulty arises because $M(\cdot)$ is not twice-differentiable at $(1/2, 0, d^0)$.

We split $M(\theta) - M(\theta_m)$ into two parts as $\{M(\theta) - M(1/2, 0, d^0)\} + \{M(1/2, 0, d^0) - M(\theta_m)\}$. Notice that by the continuity of $M(\cdot)$, the second term goes to 0. To handle the first term notice that $M(\theta) - M(1/2, 0, d^0) = \int_0^d \{\nu(x) - \alpha\}^2 p_X(x) dx + \int_d^1 \{\nu(x) - \beta\}^2 p_X(x) dx$. There exists $M_0 \in \mathbb{N}$ such that for all $m > M_0$, we have $\theta_m \in [\theta_0 - (\epsilon/2)\mathbf{1}, \theta_0 + (\epsilon/2)\mathbf{1}]$ as $\theta_m \rightarrow \theta_0$. Observe that for $\theta = (\alpha, \beta, d) \in [\theta_m - \epsilon \mathbf{1}, \theta_m + \epsilon \mathbf{1}]^c$, $m > M_0$, and $d \geq d^0$, we have

$$\begin{aligned} M(\theta) - M(1/2, 0, d^0) &= \int_0^{d^0} \left(\frac{1}{2} - \alpha \right)^2 p_X(x) dx + \int_{d^0}^d \beta^2 p_X(x) dx \\ &\quad + \int_d^1 \beta^2 p_X(x) dx \geq \int_0^{d^0} \left(\frac{1}{2} - \alpha \right)^2 p_X(x) dx \geq \left(\frac{\epsilon}{2} \right)^2 \kappa l \end{aligned}$$

as $|1/2 - \alpha| \geq |\alpha_m - \alpha| - |\alpha_m - 1/2| \geq \epsilon/2$. Similarly, for $\theta \in [\theta_m - \epsilon \mathbf{1}, \theta_m + \epsilon \mathbf{1}]^c$, $m > M_0$, and $d < d^0$, we have

$$\begin{aligned} M(\theta) - M(1/2, 0, d^0) &= \int_0^d \left(\frac{1}{2} - \alpha \right)^2 p_X(x) dx + \int_d^{d^0} \left(\frac{1}{2} - \beta \right)^2 p_X(x) dx \\ &\quad + \int_{d^0}^1 \beta^2 p_X(x) dx \geq \int_{d^0}^1 \beta^2 p_X(x) dx \geq \left(\frac{\epsilon}{2} \right)^2 \kappa l \end{aligned}$$

as $|\beta| \geq |\beta_m - \beta| - |\beta_m| \geq \epsilon/2$. Take $\eta = \kappa l \epsilon^2 / 4$. This completes the proof of the claim.

Proof of Theorem 3.2: We start with some notation. Let $W_m^{(i)} \equiv \sqrt{m} \bar{\epsilon}_{i,\cdot} / \sigma_0$, $i = 1, 2, \dots, n$, and consider our i.i.d. “data” as: $\{X_i, W_m^{(i)}\}_{i=1}^n$. Note that $W_m^{(i)}$ has density $\phi_m(\cdot)$. Let $\mathbb{P}_{n,m}(\cdot)$ denote the empirical measure of these observables and P_m the joint law of $(X_1, W_m^{(1)})$. Let σ_0 denote the true variance of ϵ_{ij} , and let σ denote any such generic value. For a fixed $\sigma > 0$ and $h \in \mathbb{R}$ define (with a slight abuse of notation):

$$\begin{aligned} Z_{im}^\sigma(h) &= 1 - \Phi \left(\frac{\sqrt{m}(\bar{Y}_i - \tau_0) - h}{\sigma} \right) = 1 - \Phi \left(\frac{\sqrt{m}(\mu(X_i) - \tau_0) - h + \sqrt{m} \bar{\epsilon}_{i,\cdot}}{\sigma} \right), \\ \mathbb{M}_{n,m}^\sigma(h) &= \frac{1}{n} \sum_{i=1}^n \left\{ Z_{im}^\sigma(h) - \frac{1}{2} \right\}^2 = \mathbb{P}_{n,m} \left[Z_{1m}^\sigma(h) - \frac{1}{2} \right]^2, \end{aligned}$$

and note that $\hat{h}_{m,n}^{\hat{\sigma}} = \arg \min_h \mathbb{M}_{n,m}^{\hat{\sigma}}(h) \equiv \sqrt{m}(\hat{\tau}_{m,n}^{\hat{\sigma}} - \tau_0)$, where $\hat{\sigma} = \hat{\sigma}_{n,m}$. Let $h_m^\sigma = \arg \min_h M_m^\sigma(h)$ where

$$\begin{aligned} M_m^\sigma(h) &= P_m \left[\frac{1}{2} - \Phi \left(\frac{\sqrt{m}(\mu(X_1) - \tau_0) - h + \sigma_0 W_m^{(1)}}{\sigma} \right) \right]^2 \\ &= \int_0^1 \left[\int_{-\infty}^{\infty} \left\{ \frac{1}{2} - g_m^{\sigma,h}(x, y) \right\}^2 \phi_m(y) dy \right] p_X(x) dx \end{aligned}$$

with $g_m^{\sigma,h}(x, y) = \Phi \left(\frac{\sqrt{m}(\mu(x) - \tau_0) - h + \sigma_0 y}{\sigma} \right)$.

Let $\epsilon, \xi > 0$ be given. Letting P_m^n denote the n -fold product measure of P_m , we want to show that $P_m^n \{|\hat{h}_{m,n}^{\hat{\sigma}} - 0| > \epsilon\} \leq \xi$ for all large m and n . Subsequently, we will denote P_m^n by P_m (again in an abuse of notation), but it will be clear from the context whether we are alluding to the product measure. We bound the quantity of interest as

$$P_m \{|\hat{h}_{m,n}^{\hat{\sigma}} - 0| > \epsilon\} \leq P_m \{|\hat{h}_{m,n}^{\hat{\sigma}} - h_m^{\hat{\sigma}}| > \epsilon/2\} + P_m \{|h_m^{\hat{\sigma}} - 0| > \epsilon/2\}. \quad (16)$$

We employ the following steps to complete the proof of the theorem:

Step 1: Establish that there exists $\delta_0 > 0$ and $M_0 > 0$ such that $|\sigma - \sigma_0| \leq \delta_0$ and $m \geq M_0$ implies $|h_m^\sigma - 0| < \epsilon/2$.

Notice that as $\hat{\sigma}$ is a consistent estimator of σ_0 , there exists M_1 such that for all $m, n \geq M_1 > 0$, $P_m \{|\hat{\sigma}_{m,n} - \sigma_0| \leq \delta_0\} \geq 1 - \xi/3$. Therefore, using *Step 1*, $P_m \{|h_m^{\hat{\sigma}} - 0| > \epsilon/2\} \leq \xi/3$ for $m \geq \max\{M_0, M_1\}$.

Step 2: Note that the first term on the right side in (16) is bounded by

$$\begin{aligned} &P_m \{|\hat{h}_{m,n}^{\hat{\sigma}} - h_m^{\hat{\sigma}}| > \epsilon/2, |\hat{\sigma} - \sigma_0| \leq \delta_0\} + P_m \{|\hat{\sigma} - \sigma_0| > \delta_0\} \\ &\leq P_m \left\{ \sup_{|\sigma - \sigma_0| \leq \delta_0} |\hat{h}_{m,n}^\sigma - h_m^\sigma| > \epsilon/2 \right\} + \xi/3 \end{aligned} \quad (17)$$

for all $n, m \geq \max\{M_0, M_1\}$. Therefore, it is enough to show that for some M (possibly depending on ϵ),

$$\sup_{m \geq M} P_m \left\{ \sup_{|\sigma - \sigma_0| \leq \delta_0} |\hat{h}_{m,n}^\sigma - h_m^\sigma| > \epsilon/2 \right\} \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (18)$$

Proof of Step 1: We study the behavior of $M_m^\sigma(h)$ as $m \rightarrow \infty$. Note that $g_m^{\sigma,h}(x, y) \rightarrow \Phi(\frac{-h + \sigma_0 y}{\sigma})$, if $x \leq d^0$, and 1 if $x > d^0$, as $m \rightarrow \infty$. Therefore, $M_m^\sigma(h)$ converges point-wise (by an application of the dominated convergence theorem (DCT) along with Scheffe's theorem) to $M^\sigma(h)$, where

$$M^\sigma(h) = c_1^\sigma(h) \int_0^{d^0} p_X(x) dx + \frac{1}{4} \int_{d^0}^1 p_X(x) dx < 1/4. \quad (19)$$

where $c_1^\sigma(h) = \int_{-\infty}^\infty \{1/2 - \Phi((-h + \sigma_0 y)/\sigma)\}^2 \phi(y) dy$. To see this, observe that $\int_{-\infty}^\infty \{1/2 - g_m^{\sigma,h}(x, y)\}^2 \phi_m(y) dy$, which is uniformly bounded by a positive constant for all m, x , can be decomposed as

$$\int_{-\infty}^\infty \{1/2 - g_m^{\sigma,h}(x, y)\}^2 \phi(y) dy + \int_{-\infty}^\infty \{1/2 - g_m^{\sigma,h}(x, y)\}^2 \{\phi_m(y) - \phi(y)\} dy \quad (20)$$

where the first term converges to $c_1(h)$ for $x \leq d^0$ and to $1/4$ for $x > d^0$. The second term in (20) converges to 0 by Scheffe's theorem for all $x \in [0, 1]$. The convergence of $M_m^\sigma(h)$ now directly follows from the DCT. Let $h^\sigma = \arg \min M^\sigma(h)$ for $h \in \mathbb{R}$.

Claim 1: There exists $\delta' > 0$, such that $\sup_{|\sigma - \sigma_0| \leq \delta'} \sup_{h \in \mathbb{R}} |M_m^\sigma(h) - M^\sigma(h)| \rightarrow 0$ as $m \rightarrow \infty$.

Proof of Claim 1: For notational convenience we will write $\Phi(x)(1 - \Phi(x))$ as $\Phi(1 - \Phi)(x)$ in what follows. Choose δ' such that $0 < \delta' < \sigma_0$. Let $\eta > 0$ be given. Note that

$$\begin{aligned} M_m^\sigma(h) - M^\sigma(h) &= A_m^{\sigma,h} \int_0^{d^0} p_X(x) dx + \int_{d^0}^1 B_m^{\sigma,h}(x) p_X(x) dx \\ \text{where } A_m^{\sigma,h} &= \int_{-\infty}^\infty \left\{ \frac{1}{2} - \Phi\left(\frac{-h + \sigma_0 y}{\sigma}\right) \right\}^2 (\phi_m - \phi)(y) dy \quad \text{and} \\ B_m^{\sigma,h}(x) &= \int_{-\infty}^\infty \left\{ \frac{1}{2} - g_m^{\sigma,h}(x, y) \right\}^2 \phi_m(y) dy - \frac{1}{4}. \end{aligned} \quad (21)$$

To simplify notation, denote the set $\{(\sigma, h) : |\sigma - \sigma_0| \leq \delta', h \in \mathbb{R}\}$ by \mathcal{C} . Then,

$$\sup_{\mathcal{C}} |M_m^\sigma(h) - M^\sigma(h)| \leq F_X(d^0) \sup_{\mathcal{C}} |A_m^{\sigma,h}| + \sup_{\mathcal{C}} \int_{d^0}^1 |B_m^{\sigma,h}(x)| p_X(x) dx.$$

Now, $\sup_{\mathcal{C}} |A_m^{\sigma,h}| \leq \int_{-\infty}^\infty |\phi_m - \phi|(y) dy \rightarrow 0$ by Scheffe's theorem, and

$$\begin{aligned} |B_m^{\sigma,h}(x)| &\leq \int_{-\infty}^\infty |\phi_m - \phi|(y) dy + \sup_{\mathcal{C}} \left| \int_{-\infty}^\infty \left\{ \frac{1}{2} - g_m^{\sigma,h}(x, y) \right\}^2 \phi(y) dy - \frac{1}{4} \right| \\ &= o(1) + \sup_{\mathcal{C}} \int_{-\infty}^\infty \Phi(1 - \Phi)\left(\frac{\sqrt{m}(\mu(x) - \tau_0) - h + \sigma_0 y}{\sigma}\right) \phi(y) dy. \end{aligned}$$

Now,

$$\sup_{\mathcal{C}} \int_{d^0}^1 |B_m^{\sigma,h}(x)| p_X(x) dx = \left(\sup_{\mathcal{C}_{\leq 0}} \int_{d^0}^1 |B_m^{\sigma,h}(x)| p_X(x) dx \right) \vee \left(\sup_{\mathcal{C}_{> 0}} \int_{d^0}^1 |B_m^{\sigma,h}(x)| p_X(x) dx \right),$$

where $\mathcal{C}_{\leq 0}$ and $\mathcal{C}_{> 0}$ are defined analogously to \mathcal{C} , but with h varying over $(-\infty, 0]$ and $(0, \infty)$, respectively. Since, for each $x > d^0$, $\sup_{\mathcal{C}_{\leq 0}} \int_{-\infty}^{\infty} \Phi(1 - \Phi)\left(\frac{\sqrt{m}(\mu(x) - \tau_0) - h + \sigma_0 y}{\sigma}\right) \phi(y) dy$ is easily seen to be dominated by $\sup_{|\sigma - \sigma_0| \leq \delta'} \int_{-\infty}^{\infty} (1 - \Phi)\left(\frac{\sqrt{m}(\mu(x) - \tau_0) + \sigma_0 y}{\sigma}\right) \phi(y) dy$ which goes to 0 as $m \rightarrow \infty$, it follows readily that the first term on the right side of the last display is $o(1)$. It remains to deal with the second. To this end, for $\lambda, h > 0$, define $D_m^{\lambda, h} = \{d^0 < x \leq 1 : |\mu(x) - (\tau_0 + h/\sqrt{m})| \leq \lambda\}$. Given $\eta > 0$, there exists $\lambda \equiv \lambda(\eta) > 0$ (but not depending on $h > 0$) such that $\int_{D_m^{\lambda, h}} p_X(x) dx < \eta$ by Assumption (A) of Theorem 3.2. Then,

$$\begin{aligned} \sup_{\mathcal{C}_{> 0}} \left| \int_{d^0}^1 B_m^{\sigma, h}(x) p_X(x) dx \right| &\leq \sup_{\mathcal{C}} \left| \int_{D_m^{\lambda, h}} B_m^{\sigma, h}(x) p_X(x) dx \right| + \sup_{\mathcal{C}_{> 0}} \left| \int_{[d^0, 1] - D_m^{\lambda, h}} B_m^{\sigma, h}(x) p_X(x) dx \right| \\ &\leq \eta + o(1) + \sup_{\mathcal{C}_{> 0}} \int_{[d^0, 1] - D_m^{\lambda, h}} \int_{-\infty}^{\infty} \Phi(1 - \Phi)\left(\frac{\sqrt{m}(\mu(x) - \tau_0 - hm^{-1/2}) + \sigma_0 y}{\sigma}\right) \phi(y) dy p_X(x) dx. \end{aligned}$$

The last term in the above display is readily seen to be bounded by

$$\sup_{\mathcal{C}_{> 0}} \int_{[d^0, 1] - D_m^{\lambda, h}} \int_{-\infty}^{\infty} \max \left\{ \Phi\left(\frac{-\sqrt{m}\lambda + \sigma_0 y}{\sigma}\right), (1 - \Phi)\left(\frac{\sqrt{m}\lambda + \sigma_0 y}{\sigma}\right) \right\} \phi(y) dy p_X(x) dx$$

which, in turn, is no larger than

$$\int_{[d^0, 1]} \int_{-\infty}^{\infty} \sup_{\sigma \in [\sigma_0 - \delta', \sigma_0 + \delta']} \max \left\{ \Phi\left(\frac{-\sqrt{m}\lambda + \sigma_0 y}{\sigma}\right), (1 - \Phi)\left(\frac{\sqrt{m}\lambda + \sigma_0 y}{\sigma}\right) \right\} \phi(y) dy p_X(x) dx$$

and this can be made less than η for sufficiently large m . It follows that

$\sup_{\mathcal{C}_{> 0}} \int_{d^0}^1 |B_m^{\sigma, h}(x)| p_X(x) dx < 3\eta$ for all sufficiently large m and Claim 1 follows.

Claim 2: There exists there exists $\delta_0 > 0$ and $M_0 > 0$ such that for all σ with $|\sigma - \sigma_0| \leq \delta_0$ and $m \geq M_0$, $|h_m^\sigma - 0| < \epsilon/2$.

Proof of Claim 2: This will be proved by a direct application of Theorem 7.1. In that theorem, take n to be m , \mathcal{T} to be the set $|\sigma - \sigma_0| \leq \delta'$ and \mathcal{H} to be \mathbb{R} . Also, \mathbb{M}_n^τ is now M_m^τ and M^τ is now M^σ . We will show that M^σ is uniquely minimized at a point, say h^σ , and also that $\inf_{|\sigma - \sigma_0| \leq \delta'} \inf_{|h - h^\sigma| > \eta} (M^\sigma(h) - M^\sigma(h^\sigma)) > 0$ for every $\eta > 0$, whence, by Claim 1, it will follow that $\sup_{|\sigma - \sigma_0| \leq \delta'} |h_m^\sigma - h^\sigma|$ converges to 0 with increasing m . But, as will also be seen, h^σ equals 0 for all σ and hence Claim 2 follows with δ_0 taken to be δ' .

From the form of $M^\sigma(h)$ (see 19) it suffices to show that $\inf_{|\sigma - \sigma_0| \leq \delta'} \inf_{|h - h^\sigma| > \eta} (c_1^\sigma(h) - c_1^\sigma(h^\sigma)) > 0$, where h^σ is the unique point at which c_1^σ is minimized. We now make some change of variables to facilitate the ensuing argument. Define $\lambda = \sigma/\sigma_0$ and $s = h/\sigma_0$. Then $|\sigma - \sigma_0| \leq \delta' \Leftrightarrow |\lambda - 1| \leq \delta''$ (for some $\delta'' < 1$) and $\Phi((-h + \sigma_0 y)/\sigma) = \Phi(\lambda^{-1}(y - s))$. Defining

$$\tilde{c}_1^\lambda(s) = \int_{-\infty}^{\infty} \left[\frac{1}{2} - \phi(\lambda^{-1}(y - s)) \right]^2 \phi(y) dy,$$

it suffices to show that $\inf_{|\lambda - 1| \leq \delta''} \inf_{|s - s^\lambda| \geq \eta/\sigma_0} (\tilde{c}_1^\lambda(s) - \tilde{c}_1^\lambda(s^\lambda)) > 0$ where s^λ is the unique minimizer of \tilde{c}_1^λ . It is easy to see that $\tilde{c}_1^\lambda(s) = E \left[\frac{1}{2} - \Phi(\lambda^{-1}(Z - s)) \right]^2$ where Z is a standard normal random variable. By the symmetry of Z about 0, it follows easily that $\tilde{c}_1^\lambda(s) = \tilde{c}_1^\lambda(-s)$. Furthermore $\tilde{c}_1^\lambda(s)$ is strictly increasing for $s > 0$, and is therefore strictly decreasing for $s \leq 0$, showing that 0 is the unique minimizer of \tilde{c}_1^λ . Hence $s^\lambda = 0$ for all λ , showing that $h^\sigma = 0$ for all σ . Thus,

$$\inf_{|\lambda - 1| \leq \delta''} \inf_{|s - s^\lambda| \geq \eta/\sigma_0} (\tilde{c}_1^\lambda(s) - \tilde{c}_1^\lambda(s^\lambda)) = \inf_{|\lambda - 1| \leq \delta''} (\tilde{c}_1^\lambda(\eta/\sigma_0) - \tilde{c}_1^\lambda(0)).$$

Since $\tilde{c}_1^\lambda(\eta/\sigma_0) - \tilde{c}_1^\lambda(0)$ is continuous and positive for each λ , its infimum on the set $|\lambda - 1| \leq \delta''$, which must be achieved, is positive.

Proof of Step 2: Consider the class of functions $\mathcal{F}_\infty \equiv \cup_m \mathcal{F}_m$ where $\mathcal{F}_m \equiv \{f_{h,\sigma}(x, w) \equiv [1/2 - \Phi(\sqrt{m}(\mu(x) - \tau_0)/\sigma + h/\sigma + w\frac{\sigma_0}{\sigma})]^2 | \tau \in \mathbb{R}, \sigma \in [\sigma_0 - \delta_0, \sigma_0 + \delta_0]\}$. This is a subclass of the large class of functions $\mathcal{G} = \{g_{\alpha,\beta,\gamma}(x, w) \equiv [1/2 - \Phi(\alpha\mu(x) + \beta w + \gamma)]^2 | (\alpha, \beta, \gamma) \in \mathbb{R}^3\}$. Note that the class $\{\alpha\mu(x) + \beta w + \gamma\}$ as (α, β, γ) varies in \mathbb{R}^3 forms a finite dimensional vector space of measurable functions and is therefore VC. Hence, $1/2 - \Phi(\alpha\mu(x) + \beta w + \gamma)$, being a bounded monotone transformation of a VC class of functions, is bounded VC and consequently, so is the class \mathcal{F}_∞ . Thus, \mathcal{F}_∞ satisfies the entropy condition in the third display on Page 168 of van der Vaart and Wellner (1996) and therefore the conditions of Theorem 2.8.1 of van der Vaart and Wellner (1996) and is uniformly Glivenko–Cantelli for the class of probability measures $\{P_m\}$, i.e., for any given $\zeta > 0$,

$$\sup_{m \geq 1} P_m \left\{ \sup_{k \geq n} \|\mathbb{M}_{m,k}^\sigma - M_m^\sigma\|_{\mathcal{F}_\infty > \zeta} \right\} \rightarrow 0 \text{ as } n \rightarrow \infty$$

and therefore,

$$\sup_{m \geq 1} P_m \left\{ \sup_{k \geq n} \|\mathbb{M}_{m,k}^\sigma - M_m^\sigma\|_{\mathcal{F}_m} > \zeta \right\} \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (22)$$

Next, using techniques similar to that from proving (15), we can show that

$$\sup_{|\sigma - \sigma_0| \leq \delta_0} |\hat{h}_{m,n}^\sigma - h_m^\sigma| > \epsilon/2 \Rightarrow \|\mathbb{M}_{m,n}^\sigma - M_m^\sigma\|_{\mathcal{F}_m} \geq \eta_m(\epsilon/2) \quad (23)$$

where $\eta_m(\epsilon) = \inf_{|\sigma - \sigma_0| \leq \delta_0} \inf_{|h - h_m^\sigma| > \epsilon/2} \{M_m^\sigma(h) - M_m^\sigma(h_m^\sigma)\}$.

Claim 3: There exists $\eta > 0$ and an integer \tilde{M} such that $\eta_m(\epsilon) \geq \eta > 0$ for all $m \geq \tilde{M}$.

Proof of Claim 3: By Claim 2, for all sufficiently large m , uniformly for $\sigma \in [\sigma_0 - \delta_0, \sigma_0 + \delta_0]$, we have $[h_m^\sigma - \epsilon/2, h_m^\sigma + \epsilon/2]^c \subset [-\epsilon/4, \epsilon/4]^c$. We conclude, that for all sufficiently large m ,

$$\eta_m(\epsilon) \geq \tilde{\eta}_m(\epsilon) \equiv \inf_{|\sigma - \sigma_0| \leq \delta_0} \inf_{|h - 0| > \epsilon/4} \{M_m^\sigma(h) - M_m^\sigma(h_m^\sigma)\}.$$

For h and σ such that $|h - 0| > \epsilon/4$ and $|\sigma - \sigma_0| \leq \delta_0$, we can bound $M_m^\sigma(h) - M_m^\sigma(h_m^\sigma)$ below as

$$\begin{aligned} M_m^\sigma(h) - M_m^\sigma(h_m^\sigma) &= (M_m^\sigma - M^\sigma)(h) - M_m^\sigma(h_m^\sigma) + M^\sigma(h) \\ &\geq - \sup_{|h - 0| > \epsilon/4} |(M_m^\sigma - M^\sigma)(h)| - M_m^\sigma(0) + M^\sigma(h) \\ &\geq - \sup_{|\sigma - \sigma_0| \leq \delta_0} \sup_{|h - 0| > \epsilon/4} |(M_m^\sigma - M^\sigma)(h)| \\ &\quad - \sup_{|\sigma - \sigma_0| \leq \delta_0} |(M_m^\sigma - M^\sigma)(0)| \\ &\quad + \inf_{|\sigma - \sigma_0| \leq \delta_0} \inf_{|h - 0| > \epsilon/4} [M^\sigma(h) - M^\sigma(0)] \end{aligned}$$

As $\sup_{|\sigma - \sigma_0| \leq \delta_0} \sup_{|h - 0| > \epsilon/4} |(M_m^\sigma - M^\sigma)(h)| \rightarrow 0$ and $\sup_{|\sigma - \sigma_0| \leq \delta_0} |(M_m^\sigma - M^\sigma)(0)| \rightarrow 0$ as $m \rightarrow \infty$, and $\inf_{|\sigma - \sigma_0| \leq \delta_0} \inf_{|h - 0| > \epsilon/4} [M^\sigma(h) - M^\sigma(0)] =: 2\eta > 0$, it follows that for all large m , $\tilde{\eta}_m(\epsilon) \geq \eta > 0$; therefore, for all sufficiently large m , say $m \geq \tilde{M}$, $\eta_m(\epsilon) \geq \eta > 0$. This completes the proof of the claim.

Hence, for all $m \geq \tilde{M}$,

$$\begin{aligned} \sup_{m \geq \tilde{M}} P_m \left\{ \sup_{|\sigma - \sigma_0| \leq \delta_0} |\hat{h}_{m,n}^\sigma - h_m^\sigma| > \epsilon/2 \right\} &\leq \sup_{m \geq \tilde{M}} P_m \left\{ \sup_{k \geq n} \|\mathbb{M}_{m,n}^\sigma - M_m^\sigma\|_{\mathcal{F}_m} > \eta_m(\epsilon)/2 \right\} \\ &\leq \sup_{m \geq \tilde{M}} P_m \left\{ \sup_{k \geq n} \|\mathbb{M}_{m,k}^\sigma - M_m^\sigma\|_{\mathcal{F}_m} > \eta/2 \right\} \end{aligned}$$

and (18) follows from (22). \square

Proof of Theorem 2.2: For notational simplicity, we refer to $\hat{d}_{m,n}$ and $\hat{\sigma}_{m,n}$ in this proof as \hat{d} and $\hat{\sigma}$ respectively. We borrow notation from the proof of Theorem 3.2, but note that now σ is a constant (as opposed to being a function). We seek to show that for given $\epsilon > 0$, $P_m\{|\hat{d} - d^0| > \epsilon\} \rightarrow 0$, as $m, n \rightarrow \infty$. Define:

$$m_d^\sigma(Z_{1m}, X_1) = \{Z_{1m}^\sigma(\tau_0) - 1/2\}^2 1(X_1 \leq d) + \{Z_{1m}^\sigma(\tau_0) - 0\}^2 1(X_1 > d),$$

and let $M_m^\sigma(d) = P_m[m_d^\sigma(Z_{1m}, X_1)]$ and $\mathbb{M}_{m,n}^\sigma(d) = \mathbb{P}_{n,m}[m_d^\sigma(Z_{1m}, X_1)]$. Also, define $d_\sigma^m = \arg \min_d M_m^\sigma(d)$ and $\hat{d}_\sigma^{m,n} = \arg \min_d \mathbb{M}_{m,n}^\sigma(d)$.

Step 0: Establish that for m sufficiently large, $\sup_{|\sigma - \sigma_0| \leq \delta_0} |d_\sigma^m - d^0| \leq \epsilon$, for a pre-assigned $\epsilon > 0$, where $\delta_0 > 0$ is a small number chosen to depend on ϵ .

This quantity of interest is dominated by:

$$P_m\{|\hat{d} - d_\sigma^m| > \epsilon/2\} + P_m\{|d_\sigma^m - d^0| > \epsilon/2\}, \quad (24)$$

where $\hat{d} \equiv \hat{d}_\sigma^{m,n}$. By Step 0 and the consistency of $\hat{\sigma}$ for σ , the second term in the above display can be made (arbitrarily) small for all sufficiently large m . By the consistency of $\hat{\sigma}$, again, it suffices to show that for some M and $\delta_0 > 0$ chosen appropriately,

$$\sup_{m \geq M} P_m \left\{ \sup_{|\sigma - \sigma_0| \leq \delta_0} |\hat{d}_\sigma^{m,n} - d_\sigma^m| > \epsilon/2 \right\} \rightarrow 0, \text{ as } n \rightarrow \infty.$$

The above display is comparable to (18) in the proof of Theorem 3.2 and follows via arguments similar (in fact, much simpler) to those following (18), involving uniformly Glivenko–Cantelli classes of functions. We omit the details and in the remainder of the proof, focus on establishing Step 0.

Recall how $W_m^{(1)}$ was defined in the proof of Theorem 3.2. We denote the distribution of $W_m^{(1)}$ by Φ_m and note that Φ_m converges weakly to Φ . Note that

$$\begin{aligned} M_m^\sigma(d) &= \int_0^d E \left[\frac{1}{2} - \Phi \left(\frac{\mu(x) - \tau_0}{\sigma m^{-1/2}} + \frac{\sigma_0}{\sigma} W_m^{(1)} \right) \right]^2 p_X(x) dx \\ &\quad + \int_d^1 E \left[1 - \Phi \left(\frac{\mu(x) - \tau_0}{\sigma m^{-1/2}} + \frac{\sigma_0}{\sigma} W_m^{(1)} \right) \right]^2 p_X(x) dx \end{aligned} \quad (25)$$

The point-wise limit of $M_m^\sigma(d)$ as $m \rightarrow \infty$ is given by

$$M^\sigma(d) = \begin{cases} c_1 \int_0^d p_X(x) dx + c_2 \int_d^{d^0} p_X(x) dx, & \text{for } d \leq d^0, \\ c_1 \int_0^{d^0} p_X(x) dx + 1/4 \int_{d^0}^d p_X(x) dx, & \text{for } d > d^0, \end{cases} \quad (26)$$

where $c_1 \equiv \int_{-\infty}^\infty \{1/2 - \Phi(\sigma_0 y / \sigma)\}^2 \phi(y) dy < \int_{-\infty}^\infty \{1 - \Phi(\sigma_0 y / \sigma)\}^2 \phi(y) dy \equiv c_2$. We first show that

$$\sup_{|\sigma - \sigma_0| \leq \delta_0, d \in [0, 1]} |M_m^\sigma(d) - M^\sigma(d)| \rightarrow 0 \quad (27)$$

as $m \rightarrow \infty$ for some $\delta_0 > 0$ sufficiently small.

Choose δ_0 such that $\sigma_0 > \delta_0 > 0$. To show that the convergence is uniform in $d \in [0, 1]$ and $\sigma \in [\sigma_0 - \delta_0, \sigma_0 + \delta_0]$, we define $g_m(\eta, \lambda) = E(\Phi(\lambda + \eta W_m^{(1)}))$, for $\lambda \in \mathbb{R}$ and $\eta > 0$. For each fixed η , $g_m(\eta, \lambda)$ is continuous in λ , converges to 0 as $\lambda \rightarrow -\infty$, to 1 as $\lambda \rightarrow \infty$ and is strictly monotone in λ .

Also define $g(\eta, \lambda) = E(\Phi(\lambda + \eta Z))$, where Z is a standard normal random variable. Let $\epsilon > 0$ be given. By Assumption (b) we can get $\beta > 0$ such that $\int_{d^0}^{d^0+\beta} p_X(x)dx < \epsilon$. Note that by Assumption (a) of the theorem there exists $\eta > 0$ such that $\mu(x) - \tau_0 > \eta$ for all $x \geq d^0 + \beta$. Let $S_m = \Phi\left(\frac{\sigma_0}{\sigma} W_m^{(1)}\right)$ and $S = \Phi\left(\frac{\sigma_0}{\sigma} Z\right)$, where $Z \sim N(0, 1)$. After some simplification, using (25) and (26), we can bound $|M_m^\sigma(d) - M^\sigma(d)|$ for $d < d^0$ by

$$\begin{aligned} & \left\{ E(1/2 - S_m)^2 - E(1/2 - S)^2 \right\} \int_0^d p_X(x)dx + \left\{ E(1 - S_m)^2 - E(1 - S)^2 \right\} \int_d^{d^0} p_X(x)dx \\ & + \epsilon + E \left[1 - \Phi \left(\frac{\eta}{\sigma m^{-1/2}} + \frac{\sigma_0}{\sigma} W_m^{(1)} \right) \right]^2 \int_{d^0+\beta}^1 p_X(x)dx. \end{aligned}$$

We can analogously bound $|M_m^\sigma(d) - M^\sigma(d)|$ for $d > d^0$. Therefore, to show that (27) holds, we need to show that

$$\sup_{|\sigma - \sigma_0| \leq \delta_0} \left| E \left[\Phi \left(\frac{\sigma_0}{\sigma} W_m^{(1)} \right) \right] - E \left[\Phi \left(\frac{\sigma_0}{\sigma} Z \right) \right] \right| \rightarrow 0, \quad \text{and} \quad (28)$$

$$\sup_{|\sigma - \sigma_0| \leq \delta_0} \left| E \left[\Phi \left(\frac{\eta}{\sigma m^{-1/2}} + \frac{\sigma_0}{\sigma} W_m^{(1)} \right) \right] - 1 \right| \rightarrow 0, \quad (29)$$

as $m \rightarrow \infty$. To show that (28) and (29) hold, we first observe that

$$|g_m(\eta, \lambda_m) - g(\eta, \lambda)| \leq |g_m(\eta, \lambda_m) - g(\eta, \lambda_m)| + |g(\eta, \lambda_m) - g(\eta, \lambda)|$$

where $\lambda_m \rightarrow \lambda$ as $m \rightarrow \infty$, with $\lambda = 0$ or $\lambda = \infty$. The second term on the right side can be controlled by using the continuity of $g(\eta, \lambda)$ for a fixed η , and the first term can be bounded by noticing that

$$\begin{aligned} & \sup_{\lambda \in \mathbb{R}, \eta \in [1-\kappa_0, 1+\kappa_0]} |g_m(\eta, \lambda) - g(\eta, \lambda)| \\ & \leq E \left[\sup_{\lambda \in \mathbb{R}, \eta \in [1-\kappa_0, 1+\kappa_0]} \left| \Phi(\lambda + \eta W_m^{(1)}) - \Phi(\lambda + \eta Z) \right| \right] \rightarrow 0 \text{ as } m \rightarrow \infty \end{aligned}$$

via an application of the DCT, where $W_m^{(1)}$ can be assumed to converge almost surely to Z (using Skorohod embedding). This proves (27).

Note that $\arg \min_{d \in [0,1]} M^\sigma(d) = d^0$. Using techniques similar to that used in proving (15), we can show that

$$\sup_{|\sigma - \sigma_0| \leq \delta_0} |d_\sigma^m - d^0| > \epsilon \Rightarrow \sup_{|\sigma - \sigma_0| \leq \delta_0, d \in [0,1]} |M_m^\sigma(d) - M^\sigma(d)| > \eta(\epsilon)/2 \quad (30)$$

where $\eta(\epsilon) = \inf_{|\sigma - \sigma_0| \leq \delta_0} \inf_{|d_\sigma^m - d^0| > \epsilon} \{M^\sigma(d_\sigma^m) - M^\sigma(d^0)\} > 0$ (follows from (26)). But, as (27) holds, (30) cannot hold for all large m , thereby completing the proof of Step 0. \square

Proof of Theorem 3.1: Letting $Z_{im}^\sigma(\tau)$ as in (7) (in Section 3) and $m_d^{\sigma,\tau}(Z_{1m}, X_1) = \{Z_{1m}^\sigma(\tau) - 1/2\}^2 1(X_1 \leq d) + \{Z_{1m}^\sigma(\tau) - 0\}^2 1(X_1 > d)$, define

$$\mathbb{M}_{m,n}^{\sigma,\tau}(d) = \mathbb{P}_{n,m}[m_d^{\sigma,\tau}(Z_{1m}, X_1)] \quad \text{and} \quad M_m^{\sigma,\tau}(d) = P_m[m_d^{\sigma,\tau}(Z_{1m}, X_1)]. \quad (31)$$

Define $\tilde{d}_{\sigma,\tau}^{m,n} = \arg \min_{d \in [0,1]} \mathbb{M}_{m,n}^{\sigma,\tau}(d)$ and $d_{\sigma,\tau}^{m,n} = \arg \min_{d \in [0,1]} M_m^{\sigma,\tau}(d)$ and note that $\tilde{d}_{m,n} = \tilde{d}_{\tilde{\sigma}, \tilde{\tau}}^{m,n}$. Let $\epsilon > 0$ be given. We seek to show that: $P_m(|\tilde{d}_{m,n} - d^0| > \epsilon) \rightarrow 0$ as $m, n \rightarrow \infty$, to which end it suffices to show that both $P_m(|\tilde{d}_{m,n} - d_{\tilde{\sigma}, \tau_0}^m| > \epsilon/2)$ and $P_m(|d_{\tilde{\sigma}, \tau_0}^m - d^0| > \epsilon/2)$ go to 0. The second term is easily handled on noting that $\sup_{|\sigma - \sigma_0| \leq \delta_0} |d_{\sigma, \tau_0}^m - d^0| \leq \epsilon/2$ for

some appropriately chosen δ_0 (using Step 0 of Theorem 2.2) and the consistency of $\tilde{\sigma}_{m,n}$ for σ_0 . To handle the first term, we show that

$$P_m \left(\sup_{|\sigma - \sigma_0| \leq \delta_0} |\tilde{d}_{\sigma, \tilde{\tau}}^{m,n} - d_{\sigma, \tau_0}^m| > \epsilon/2 \right) \rightarrow 0,$$

and this, again in conjunction with the consistency of $\tilde{\sigma}_{m,n}$ implies the convergence of the first term to 0. We next introduce some notation. For a real-valued function x^σ defined on $[0, 1]$, define $\|x^\sigma\| = \sup_{|\sigma - \sigma_0| \leq \delta_0} \sup_{d \in [0, 1]} |x^\sigma(d)|$, where $\sigma_0 > \delta_0 > 0$. By arguments analogous to *generic steps*, we have:

$$P_m \left(\sup_{|\sigma - \sigma_0| \leq \delta_0} |\tilde{d}_{\sigma, \tilde{\tau}}^{m,n} - d_{\sigma, \tau_0}^m| > \epsilon/2 \right) \leq P_m(\|\mathbb{M}_{m,n}^{\sigma, \tilde{\tau}} - M_m^{\sigma, \tau_0}\| \geq \eta_m(\epsilon)/2),$$

with $\eta_m(\epsilon) = \inf_{|\sigma - \sigma_0| \leq \delta_0} \inf_{|d - d_{\sigma, \tau_0}^m| > \epsilon/2} \{M_m^{\sigma, \tau_0}(d) - M_m^{\sigma, \tau_0}(d_{\sigma, \tau_0}^m)\}$. By using arguments similar to that of Claim 3 of Theorem 3.2 and the Claim of Theorem 2.1 (following (15)), we can show that $\eta_m(\epsilon) > \eta > 0$ for all sufficiently large m , whence it suffices to show that $P_m(\|\mathbb{M}_{m,n}^{\sigma, \tilde{\tau}} - M_m^{\sigma, \tau_0}\| \geq \eta/2)$ goes to 0 as $m, n \rightarrow \infty$. Now,

$$\|\mathbb{M}_{m,n}^{\sigma, \tilde{\tau}} - M_m^{\sigma, \tau_0}\| \leq \|\mathbb{M}_{m,n}^{\sigma, \tilde{\tau}} - \mathbb{M}_{m,n}^{\sigma, \tau_0}\| + \|\mathbb{M}_{m,n}^{\sigma, \tau_0} - M_m^{\sigma, \tau_0}\|. \quad (32)$$

Using arguments (involving universal Glivenko-Cantelli classes) similar to the proof of Step 2 in Theorem 3.2, we can easily show that: $\sup_{m \geq 1} P_m(\sup_{k \geq n} \|\mathbb{M}_{m,k}^{\sigma, \tau_0} - M_m^{\sigma, \tau_0}\| > \eta/4) \rightarrow 0$ as $n \rightarrow \infty$, whence it readily follows that $P_m(\|\mathbb{M}_{m,n}^{\sigma, \tau_0} - M_m^{\sigma, \tau_0}\| > \eta/4) \rightarrow 0$ as $m, n \rightarrow \infty$. That $P_m(\|\mathbb{M}_{m,n}^{\sigma, \tilde{\tau}} - \mathbb{M}_{m,n}^{\sigma, \tau_0}\| > \eta/4)$ goes to 0 follows from the fact that $\|\mathbb{M}_{m,n}^{\sigma, \tilde{\tau}} - \mathbb{M}_{m,n}^{\sigma, \tau_0}\| \leq \tilde{K} \sqrt{m}(\tilde{\tau}_{m,n} - \tau_0)$ which goes to 0 by assumption, as $m, n \rightarrow \infty$. This last inequality follows from the fact that for any $d \in [0, 1]$:

$$\begin{aligned} |\mathbb{M}_{m,n}^{\sigma, \tilde{\tau}}(d) - \mathbb{M}_{m,n}^{\sigma, \tau_0}(d)| &\leq \frac{1}{n} \sum_{i: X_i \leq d} 3|Z_{im}^\sigma(\tilde{\tau}) - Z_{im}^\sigma(\tau_0)| + \frac{1}{n} \sum_{i: X_i > d} 2|Z_{im}^\sigma(\tilde{\tau}) - Z_{im}^\sigma(\tau_0)| \\ &\leq \frac{3}{n} \sum_{i=1}^n |Z_{im}^\sigma(\tilde{\tau}) - Z_{im}^\sigma(\tau_0)| \leq \frac{\sqrt{m}(\tilde{\tau} - \tau_0)}{\sigma} \frac{1}{n} \sum_{i=1}^n \phi(\xi_i^*)/\sigma. \\ &\leq K \frac{\sqrt{m}(\tilde{\tau} - \tau_0)}{\sigma} \xrightarrow{p} 0 \end{aligned} \quad (33)$$

for some universal constant K , whence \tilde{K} can be taken to be $K/(\sigma_0 - \delta_0)$. \square .

Justification of the subsampling procedure: We consider an asymptotic paradigm, where m is viewed as fixed, and n as increasing to infinity in the setting of a *known* τ_0 (which can then be taken to be 0 without loss of generality). Let the setting be that of Theorem 2.1 and consider fitting a stump $\xi_d(x) = (1/2)1(x \leq d)$ as a *working model* for ν_m . Recall that the underlying setting corresponds to a regression one with i.i.d. observations $\{X_i, Z_{im}\}_{i=1}^n$. Letting P_m denote the distribution of $(X_1, Z_{1,m})$, the best-fitting stump in the population is characterized by the parameter $d^m := \arg \min M_m(d)$, where $M_m(d) = P_m[(Z_{1,m} - (1/2))^2 1(X_1 \leq d) + Z_{1,m}^2 1(X_1 > d)]$. Setting the derivative of M_m with respect to d , to zero, yields the normal equation $\nu_m(d^m) = 1/4$. Under reasonably modest assumptions on the underlying model, d^m is unique and provides an upper bound on d^0 and the larger the m , the tighter the bound. A level $1 - \alpha$ confidence interval for d^m can then be used as a *surrogate* for a level $1 - \alpha$ confidence interval for d^0 . Letting $\hat{d}_{m,n}$ be the least squares estimate of d^0 obtained by minimizing $\sum_{i=1}^n [(Z_{i,m} - 0.5)^2 1(X_i \leq d) + Z_{i,m}^2 1(X_i > d)]$ over all d , it can be shown, by adapting the techniques of Banerjee and McKeague (2007), that for a fixed m , $n^{1/3}(\hat{d}_{m,n} - d^m)$ converges to a continuous symmetric but non-Gaussian distribution, namely Chernoff's distribution (studied in Groeneboom and Wellner (2001)),

as $n \rightarrow \infty$. As this distribution depends upon hard to estimate nuisance parameters in the model, subsampling without replacement (as in Politis, Romano and Wolf (1999)) or the m out of n bootstrap can be used to extract asymptotic confidence intervals for d^m . Owing to the non-standard nature of the asymptotics involved, the standard Efron-type bootstrap fails in this situation (see Sen, Banerjee and Woodroffe (2010) and references therein).

Note that the implemented subsampling procedure, while relying on the above results in spirit, does take some liberties in its implementation. Firstly, the Method 1 based confidence intervals require estimation of τ since it is unknown for our application. Secondly, the theoretical results above are not immediately applicable to the confidence intervals using Method 2 which involves a non-trivial modification of the p -values used in Method 1. Nevertheless, it seems reasonable to conjecture the same $n^{1/3}$ rate of convergence to d^m for the least squares estimates obtained by this approach, with a non-degenerate continuous distribution, which would then validate the use of subsampling.

References

- [1] Bambos, N. and Michailidis, G. (2004) Queueing and Scheduling in Random Environments. *Adv. Appl. Prob.*, **36**, 293–317.
- [2] Banerjee, M. and McKeague, I. (2007) Confidence Sets for Split Points in Decision Trees. *Ann. Statist.*, **35**, 543–574.
- [3] Banerjee, M. and Wellner, J. A. (2005) Confidence intervals for current status data. *Scand. J. Statist.*, **32**, 405–424.
- [4] Chen, Y. (1999) Nonparametric Identification of the Minimum Effective Dose. *Biometrics*, **55**, 1236–1240.
- [5] Chen, Y. and Chang, Y. (2007) Identification of the minimum effective dose for right-censored survival data. *Comp. Statist. Data Ana.*, **51**, 3213–3222.
- [6] Cox, C. (1987) Threshold dose-response models in toxicology. *Biometrics*, **43**, 511–523.
- [7] Feder, P. I. (1975) On asymptotic distribution theory in segmented regression problems — identified case. *Ann. Statist.*, **3**, 49–83.
- [8] Groeneboom, P. and Wellner, J. A. (2001) Computing Chernoff’s distribution. *J. Comput. Graph. Statist.*, **10**, 388–400.
- [9] Gruet, M.A. (1996) A nonparametric calibration analysis. *Ann. Statist.*, **24**, 1474–1492.
- [10] Hsu, J. and Berger, R. (1999) Stepwise confidence intervals without multiplicity adjustment for dose-response and toxicity studies. *J. Amer. Statist. Assoc.*, **94**, 468–482.
- [11] Keshamouni, V. G., Michailidis, G., Grasso, C. et al. (2006), Differential protein expression profiling by iTRAQ 2D-LC-MS/MS of lung cancer cells undergoing epithelial-mesenchymal transition reveals a migratory/invasive phenotype, *J. Proteome Res.*, **5**, 1143–1154.
- [12] Keshamouni, V. G., Jagtap, P., Michailidis, G., et al. (2009), Temporal quantitative proteomics by iTRAQ 2D-LC-MS/MS and corresponding mRNA Expression analysis identify post-transcriptional modulation of actin-cytoskeleton regulators during TGF- β -induced epithelial-mesenchymal transition, *J. Proteome Res.*, **8**, 35–47.
- [13] Koul, H. L. and Qian, L. (2002) Asymptotics of maximum likelihood estimator in a two-phase linear regression model. *C. R. Rao 80th birthday felicitation volume, Part II. J. Statist. Plann. Infer.*, **108**, 99–119.
- [14] Lan, Y., Banerjee, M. and Michailidis, G. (2009) Change-point estimation under adaptive sampling, *Ann. of Statist.*, **37**, 1752–1791.

- [15] Loader, C. R. (1996) Change point estimation using nonparametric regression. *Ann. of Statist.*, **24**, 1667–1678.
- [16] Mueller, H. G. (1992) Change-points in nonparametric regression analysis. *Ann. Statist.*, **20**, 737–761.
- [17] Mueller, H. G. and Song, K. S. (1997) Two-stage change-point estimators in smooth regression models. *Statist. Probab. Lett.*, **34**, 323–335.
- [18] Osborne, C. (1991) Statistical calibration: A review. *Intl. Stat. Rev.*, **59**, 309–336.
- [19] Politis, D. N., Romano, J. P. and Wolf, M. (1999) *Subsampling*. Springer-Verlag, New York.
- [20] Raimondo, M. (1998) Minimax estimation of sharp change points. *Ann. Statist.*, **26**, 1379–1397.
- [21] Robertson, T., Wright, F. T. and Dykstra, R.L. (1988) *Order restricted statistical inference*. Wiley, New York.
- [22] Silvapulle, M. J. and Sen, P. K. (2005) *Constrained statistical inference. Inequality, order, and shape restrictions*. Wiley Series in Probability and Statistics.
- [23] Sen, B., Banerjee, M. and Woodroffe, M. (2010) Inconsistency of Bootstrap: The Grenander estimator. To appear in *Ann. Statist.*
- [24] Sen, B., Banerjee, M., Woodroffe, M., Mateo, M. and Walker, M. (2009) Streaming motion in Leo I. *Ann. Appl. Statist.*, **3**, 96–116.
- [25] Singh, A., Scott, C. and Nowak, N. (2009) Adaptive Hausdorff estimation of density level sets. *Ann. Statist.*, **37**, 2760–2782.
- [26] Tamhane, A. and Logan, B. (2002) Multiple test procedures for identifying the minimum effective and maximum safe doses of a drug. *J. Amer. Statist. Assoc.*, **97**, 293–301.
- [27] Tang, R., Banerjee, M. and Michailidis, G. (2010) A two-stage hybrid procedure for estimating an inverse regression function. To appear in *Ann. Statist.*
- [28] Van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes*. Springer, New York.
- [29] Willet, R. and Nowak, R. (2007) Minimax optimal level set estimation. *IEEE Trans. Inform. Theory*, **16**, 2965–2979.
- [30] Williams, D. A. (1971) A Test for Differences between Treatment Means When Several Dose Levels are Compared with a Zero Dose Control. *Biometrics*, **27**, 103–117.